COMPUTATIONAL AND EXPERIMENTAL INVESTIGATION OF THE ENZYMATIC HYDROLYSIS OF CELLULOSE

A Dissertation
Presented to
The Academic Faculty

by

Prabuddha Bansal

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Chemical & Biomolecular Engineering

Georgia Institute of Technology
December 2011

COMPUTATIONAL AND EXPERIMENTAL INVESTIGATION OF THE ENZYMATIC HYDROLYSIS OF CELLULOSE

Approved by:

Dr. Matthew J. Realff, Advisor
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

Dr. Mark P. Styczynski
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

Dr. Andreas S. Bommarius, Advisor
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

Dr. Joshua S. Weitz
School of Biology
Georgia Institute of Technology

Dr. Jay H. Lee, Advisor
Department of Chemical and Biomolecular Engineering, *Korea Advanced Institute of Science and Technology (KAIST)*, and School of Chemical & Biomolecular Engineering, *Georgia Institute of Technology*

Dr. Ronald W. Smith
Chevron Energy Technology Company

Date Approved: August 24, 2011

If you're not prepared to be wrong, you will never come up with anything original

– Sir Kenneth Robinson

To late Mrs. Shanta Bansal, late Mr. Miri Mal Bansal, Mrs. Anika Sur, late Mr. Arun Kumar Sur, and to the undying love of my family.

ACKNOWLEDGMENTS

There are many people who have directly or indirectly been helpful in both, the successful completion of this thesis, as well as making my years in Atlanta and Georgia Tech memorable. Though the few words in this section may not be able to convey my full gratitude to them, I will try my best to do so.

First and foremost, I would like to thank my advisors (in alphabetical order of first name): Dr. Andreas S. Bommarius, Dr. Jay H. Lee, and Dr. Matthew J. Realff. Working with three advisors, which seemed challenging at first, actually turned out to be very smooth, and fruitful too. The combination of their knowledge on various fields such as protein engineering, biocatalysis, systems engineering, and machine learning, gave me new insights into how protein engineering and the enzymatic hydrolysis of cellulose can be studied. In my years as a PhD student, I have also developed a lot professionally, and I am indebted to my advisors for that. I also thank them for believing in me, and advising me whenever I hit road blocks.

I would like to thank my committee members, Dr. Ron Smith, Dr Mark P. Styczynski, and Dr. Joshua S. Weitz, for their constant input and review of my thesis work. Dr. Ron Smith and Dr. Jay H. Lee, who are present in California and South Korea respectively, have been very adjusting in agreeing to telephone conference at a time inconvenient to them. I would like to thank them for this.

Without Dr. Mélanie Hall, much of the work in this thesis would not have been possible. With her I have co-authored four publications so far, and will be submitting a fifth one in the near future. Not only did she contribute in terms of experimental data (sometimes on a short notice), she also guided me through the writing of my first paper,

and engaged in many productive research discussions. I am very grateful to her for all the help.

I would also like to thank Bryan J. Vowell, an undergraduate research student in the Bommarius lab, who was very instrumental in acquiring data with the cellulose hydrolysis and adsorption experiments, sometimes going above and beyond the call of duty. He was also patient enough to bear with me when I was getting trained on the experimental procedures. I also thank Yuzhi Kang, for helping with different assays and contributing ideas. I would also like to extend my thanks to Dr. Yanto Yanto, and Jonathan Park for collaborating on protein engineering of Old Yellow Enzymes.

I extend my thanks to all the members of the Bommarius and Lee group. With Dr. Wee Chin Wong, Dr. Farminder Singh Anand, Dr. Ugur Guner, and Dr. Nikolaos Pratikakis, I had many fun moments and intellectual discussions, both inside and outside of school. I also thank Dr. Rakshita Agrawal and Kevin Yeh for sharing office space with me. The Bommarius group members, alumni and current members, have been excellent colleagues and friends: Dr. Mélanie Hall, Dr. Thomas A. Rogers, Dr. Janna K. Blum, Dr. Yanto Yanto, Dr. Eduardo Vazquez-Figueroa, Andria Deaguero, Michael Abrahamson, Russell Vegh, Michael K. Rood, Jonathan Rubin, Jonathan Park, Yuzhi Kang, and Ryan Clairmont.

Dr. Guhan Jayaraman, in whose lab I spent my last year at IIT Madras, guided me through my initial experiences of scientific research, and I would like to thank him for that.

I'd like to thank my family, and friends in Atlanta, USA, and India whose constant support and interaction made it possible for me to make it through some stressful

times. Without Geetika Agarwal, Ankur Gupta, Ugur, Wee Chin, Farminder, Pramod, Prashant, Siddharth, Milky, Anil, Nitin Arora, Kalyan, Salil, Aritra, Ambarish, Nitesh, Tanushree, Rohan, Divya, Manoj, and many more, my stay in Atlanta would not have been so much fun. At this moment I would also like to remember my friends from IIT Madras (batch-mates, wing-mates, seniors, and the soccer team) and Chandigarh, with whom I spent some memorable times. Finally, the constant love and encouragement of my parents and elder brother, was invaluable in the successful completion of my thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

SUMMARY

The enzymatic hydrolysis of cellulose to glucose by cellulases is one of the major steps involved in the conversion of lignocellulosic biomass to yield biofuel. This hydrolysis by cellulases, a heterogeneous reaction, currently suffers from some major limitations, most importantly a dramatic rate slowdown at high degrees of conversion in the case of crystalline cellulose. Elucidation of the major rate-limiting factors has been impeded by interaction of various substrate- and enzyme-related properties. In this thesis, computational as well as experimental studies have been pursued to extricate the causes of rate deceleration in the enzymatic hydrolysis of cellulose. To guide protein engineering on cellulases, cellulose-binding domains (CBDs), or in general on proteins for which there is not a high-throughput assay available, a novel method to suggest target mutations was developed.

An extensive review of literature on modeling enzymatic hydrolysis of cellulose has been provided (Chapter 2). Although the various hypotheses of rate-limiting factors that were employed to develop and validate these models have shed light on the possible rate hindrances, a unified picture of the mechanism of rate retardation in cellulose biohydrolysis is missing. Accumulation of works with invalid assumptions of Michaelis-Menten kinetics and quasi-steady state, empirical factors accounting for rate decline, and conflicting reports on contribution of various rate-limiting factors to the rate decline has further slowed down the understanding of the rate limitations.

Applying experimental as well as computational tools, various rate limiting factors were screened to identify substrate accessibility to cellulases and hydrolysability

(hydrolysable fraction of accessible substrate) as the major rate hindrances (Chapter 3). Reactivity, defined in terms of hydrolytic activity per amount of actively adsorbed cellulase, was observed to remain constant with conversion. Enzyme clogging was observed in the form of higher restart rates as compared to the uninterrupted rates.

Cellulose crystallinity is a major substrate property affecting the rates, but its quantification has suffered from lack of consistency and accuracy. Using multivariate statistical analysis of X-ray data from cellulose, a new method to determine the degree of crystallinity was developed (Chapter 4). Principal component analysis was also employed, to examine the high dimensional nature of the X-ray spectra. The method was successfully validated with leave-one-out validation, and with mechanically prepared cellulose samples of known crystallinity. It was also applied successfully to Cel7A (cellobiohydrolase I from *Trichoderma reesei*) CBD, and partially converted Avicel. A strong linear relationship between the degree of crystallinity and initial hydrolysis rate provided evidence of crystallinity as a major determinant of rates.

Cel7A CBD is a promising target for protein engineering as cellulose pretreated with Cel7A CBDs exhibits enhanced hydrolysis rates resulting from a reduction in crystallinity. However, for Cel7A CBD, a high throughput assay is unlikely to be developed since cellulose pretreatment requires long incubation times. In the absence of a high throughput assay (required for directed evolution) and extensive knowledge of the role of specific protein residues (required for rational protein design), the mutations need to be picked wisely, to avoid the generation of inactive variants. To tackle this issue, a method utilizing the underlying patterns in the sequences of a protein family has been developed (Chapter 5). The low dimensional topology of the sequences is identified *via*

principal component analysis, and is used to look for changes in the sequence of interest such that it is closer to the identified manifold. These changes are picked as the mutations of interest. Since activity data is not employed, it is a case of unsupervised machine learning. This method was successfully shown to identify beneficial mutations in a literature data set.

The work presented in this thesis ranges from kinetic studies and statistical methods for studying cellulose biohydrolysis to applying data-mining tools for protein engineering, and is a good example of how heterogeneous biocatalysis can be improved. Finally, the conclusions are presented in the last chapter (Chapter 6), along with an outlook on how the work presented in thesis can be extended.

CHAPTER 1

INTRODUCTION

Increasing demand of fuel and energy, concerns over greenhouse emissions, and independence from petroleum based fuels has led to a recent increase in interest in biofuels in the recent past. Lignocellulose, consisting of lignin, hemicelluloses and cellulose, is the largest naturally occurring carbon based material, and represents a major potential feedstock for biofuel. Ethanol derived from lignocellulose is produced *via* four major consecutive steps: pretreatment, hydrolysis, fermentation, and separation. For cellulosic ethanol to compete economically with gasoline and corn ethanol, major improvements have to be made in the enzymatic hydrolysis of cellulose (Galbe and Zacchi, 2002; Lynd et al., 2008; Sun and Cheng, 2002). According to the numbers in a recent design and economics study on biochemical conversion of lignocellulosic biomass to ethanol, enzyme contribution to the conversion cost (excluding feedstock cost) is about 25% (http://www.nrel.gov/docs/fy11osti/47764.pdf). The major limitations in enzymatic breakdown of cellulose are high cost of enzymes, and slow rates of hydrolysis as exhibited in their dramatic slowdown at high degrees of conversion (Figure 1). This is not simply due to substrate depletion (Chapter 3).

The cost contribution of cellulases per gallon of ethanol can be factored into two terms: $/gal ethanol = ($/enzyme)*(enzyme/gal ethanol). The cost contribution of cellulases can be reduced i) by improvements in cellulase expression, the focus of many biotechnology companies (Schubert, 2006), ii) by improving the cellulase machinery itself through *protein engineering* (Himmel et al., 2007), and iii) by improving the rate of cellulose hydrolysis through optimization of reaction conditions *via process engineering*, accomplished in good part by rendering the cellulose substrate less recalcitrant to enzymatic action (*substrate engineering*). Advances in these three areas depend strongly

on the quantification of the key enzyme-substrate interactions and enzyme/substrate properties determining the rates, and identification of the causes for the rate slowdown.

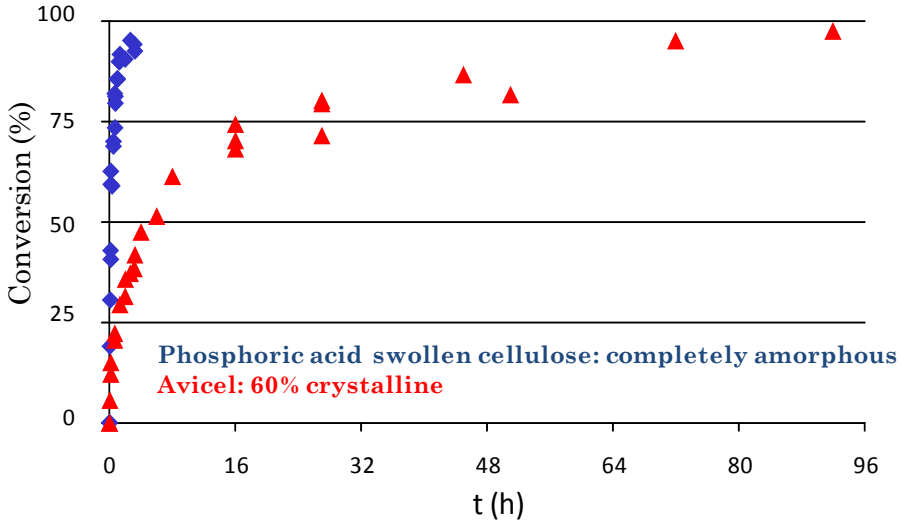


Figure 1. Conversion-time profiles of Avicel (a commercially available 60% crystalline cellulose) and phosphoric acid swollen cellulose (PASC, completely amorphous, generated by pretreating Avicel with phosphoric acid).

Cellulose is degraded synergistically into glucose by three types of cellulases: endoglucanases (EC 3.2.1.4), that randomly cleave β-1,4-glycosidic bonds on cellulose chains away from chain ends, cellobiohydrolases (EC 3.2.1.91), that produce cellobiose by attacking cellulose from chain ends (Cel7A (cellobiohydrolase I), acts from the reducing ends, and Cel6A (cellobiohydrolase II) acts from the non-reducing ends of the cellulose chains) as well as β-glucosidases (EC 3.2.1.21) that convert cellobiose to glucose (Henrissat, 1994; Lynd et al., 2002; Rabinovich et al., 2002; Teeri, 1997; Zhang and Lynd, 2004). Biohydrolysis of cellulose, due its heterogeneous nature, involves more steps than classical enzyme kinetics. The major steps are (Figure 2):

1. Adsorption of cellulases onto the substrate via the binding domain (Ståhlberg et al., 1991),

2. Location of a bond susceptible to hydrolysis on the substrate surface (Jervis et al., 1997) (chain end if cellobiohydrolase, cleavable bond if endoglucanase),

3. Formation of enzyme–substrate complex (by threading of the chain end into the catalytic tunnel if cellobiohydrolase, to initiate hydrolysis) (Divne et al., 1998; Mulakala and Reilly, 2005),

4. Hydrolysis of the β-glycosidic bond (involving cleavage of the β-glycosidic bond, cellobiose product expulsion from the active site and decrystallization) and simultaneous forward sliding of the enzyme along the cellulose chain (Beckham et al., 2011; Divne et al., 1998; Mulakala and Reilly, 2005),

5. Desorption of cellulases from the substrate or repetition of step 4 or steps 2/3 if only the catalytic domain detaches from chain,

6. Hydrolysis of cellobiose to glucose by β-glucosidase (if present in the enzyme mixture). In addition, product inhibition (Bezerra and Dias, 2005; Holtzapple et al., 1990; Xiao et al., 2004; Yue et al., 2004) and changes in the substrate properties along the course of hydrolysis affect the above steps.

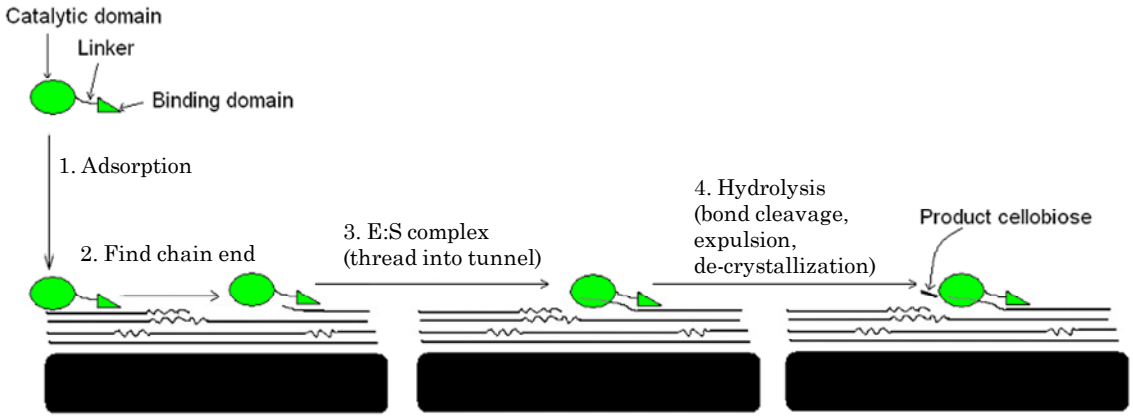


Figure 2. Steps 1 to 4 for a cellobiohydrolase acting on a cellulosic substrate (not drawn to scale). For endoglucanases, steps 2 and 3 are different as it does not require chain ends to act on.

While the above mentioned steps are known to be the major ones, modeling cellulase kinetics has not been easy. This is mainly due to the unclear factors responsible for the rapidly decelerating rates along conversion (Figure 1), imposing additional layers of complexity from a kinetic modeling perspective. Studies have pointed to various interplaying phenomena resulting in the precipitous reduction in hydrolysis rates. However, determination of the dominant factors and their exact quantification still eludes the cellulose community. Mechanistic understanding of the enzymatic action on cellulose is further impeded by the fact that many models in the literature suffer from over parameterization (Sin et al., 2009), empirical factors to account for various enzyme and substrate properties, and invalid assumptions like Michaelis-Menten and quasi steady-state conditions. More incisive kinetic studies are required to validate or refute the various rate limitation hypotheses.

The mechanism of cleavage of the β-glycosidic bond in the active site of the cellobiohydrolases and endoglucanases is well known (Schulein, 2000). Improvements in enzyme catalysis have mainly been guided by the engineering of the active site or amino acid residues identified as playing an important role. In the case of cellulases and their kinetics on insoluble lignocellulosic substrates, rate limitations cannot be explained solely by active-site considerations, mostly because of the heterogeneity of the substrate. Recently, it has been shown that pretreatment of cellulose with cellulose binding domains (CBDs) from *Trichoderma reesei* cellobiohydrolase-I (Cel7A) results in enhanced hydrolysis rates due to a reduction in crystallinity (Hall et al., 2011). This two pronged strategy of targeting the major rate determining substrate property with an agent that comes from the major cellulase component itself, has scope for further improvement through protein engineering of the Cel7A CBD.

Protein engineering for biocatalysts has developed in three major phases: first phase of rational design, second phase driven by directed evolution or combinatorial

design, and most recently, data-driven and computational protein design (Bommarius et al., 2011). Rational design is driven by structure-function relationships, where the search space is reduced by three dimensional structural considerations, improving the chances of a 'hit'. Directed evolution (Arnold and Volkov, 1999) on the other hand, requires no knowledge of the protein's three dimensional structure, and has emerged as a widely deployed tool in the last two decades. The most efficient and commonly used form of directed evolution is DNA shuffling (Crameri et al., 1998; Stemmer, 1994), where first random fragments are made from DNA templates for recombination to generate diversity, and then the DNA sequences (of protein variants) with improved function are chosen to further undergo evolution and recombination with the aim of achieving higher functionality.

Computational tools for improving directed evolution, utilizing data in form of sequence diversity, structures, and substrate specificities, have emerged recently with a lot of promise (Bommarius et al., 2011). Another recent data-driven approach involves advanced statistical tools or machine learning to help select variants for testing (Fox and Huisman, 2008). In this approach, a statistical model mapping the protein sequence space on to the functionality guides the mutation strategy. This approach becomes highly advantageous for engineering enzymes whose structural knowledge in not very well known (for rational design), or for which a high throughput assay is not available (for directed evolution). Essentially all the machine learning tools are predicated on being able to extract useful information about the sequence to function mapping from limited samples. This has been enabled by the drop in the cost and time of sequencing variants and the experimental ability to generate specific sets of mutants as opposed to relying on random generation techniques.

For Cel7A and Cel7A CBD, a directed evolution approach is unlikely to be successful due to the complexity of the function and size of the protein. This is coupled with the difficulty of establishing a high throughput assay, since cellulose pretreatment is

slow and involves a solid phase reaction with no cheap or reliable indicator of reaction progress. The rational approach is hampered by the lack of understanding of the rate limitations of cellulose degradation.

Recent works on statistical protein engineering include: ProSAR (Protein Structure Activity Relationship) using partial least squares regression (Fox et al., 2007), probabilistic modeling for protein systems (Barak et al., 2008; Brouk et al., 2010; Nov and Wein, 2005), 'Mt. Fuji' type landscape model (Aita et al., 2001; Aita et al., 2000), and usage of multiple regression methods (Liao et al., 2007). One of the major bottlenecks for almost all protein engineering methods is the generation of the initial set of variants. Even some of the above-mentioned works, for which this step serves to provide the initial data set for model building, use directed evolution to create the first set of mutants. Under conditions of low throughput assays however, this may not be feasible. In spite of all the protein engineering efforts so far, systematic methods for identifying target mutations *a priori* to any experimental work remain elusive. Examples of published techniques that have shown promise include SCHEMA (Meyer et al., 2003), Rosetta (Siegel et al., 2010), and CASTing (Reetz and Carballeira, 2007).

1.1 Research contributions

A thorough survey of the literature on modeling enzymatic hydrolysis of cellulose (Bansal et al. (2009); Chapter 2) has provided a comprehensive overview and a critical analysis of current models, their basic assumptions, and their usefulness as well as shortcomings. Modeling and simulation studies in conjunction with experiments have been pursued in this thesis to screen the various hypotheses for causing the rate slowdown, and successfully quantify the contributions of the relevant rate hindrances (Chapter 3).

Cellulose crystallinity is one of the major substrate properties determining the hydrolysis rate and has been the subject of investigation in many studies (Bansal et al.,

2009; Lynd et al., 2002; Zhang and Lynd, 2004). Although a number of methods to calculate the degree of crystallinity of cellulose from X-ray diffraction spectra have been published, different crystallinity values can be extracted using different analytical methods on the same spectrum (Park et al., 2010; Thygesen et al., 2005). To address this issue, a new data-driven method to determine the degree of crystallinity of cellulose from X-ray diffraction spectra was developed (Bansal et al. (2010); Chapter 4). This method was shown to give accurate and consistent crystallinity index values for pure cellulosic substrates, Cel7A CBD pretreated cellulose, and partially converted Avicel. It is now being extended by Yuzhi Kang in the Bommarius lab to lignocellulosic substrates.

In the absence of a high-throughput assay (required for directed evolution) and extensive knowledge of the role of specific protein residues (required for rational protein design), mutations for protein engineering need to be picked wisely, guided by some other methodology to avoid the generation of oversized libraries of mutants. A new method utilizing the underlying patterns in a protein family's sequences has been developed and tested with both literature as well as experiments (Chapter 5).

CHAPTER 2

REVIEW OF KINETIC MODELS AND RATE HINDRANCES IN THE ENZYMATIC HYDROLYSIS OF CELLULOSE

(Parts of this chapter are reproduced from Bansal et al. (2009). Text is updated with works published since then)

2.1 Introduction

Experimental data on cellulose hydrolysis by cellulases point to various bottlenecks that contribute to decreasing rates with conversion. To deconvolute the data, mathematical modeling of the hydrolysis process is an important tool. Further improvement of cellulase kinetics will be guided by the relative importance of physical parameters of the model, such as those associated with adsorption or surface accessibility. To find and alleviate bottlenecks, the kinetic and the physical parameters in the model have to be estimated correctly. Lee et al. (1980) reviewed the models published up to that point. Zhang and Lynd (2004) discussed the potential use of various models in literature, based on the number of substrate and enzyme variables considered. Both these articles concluded that to achieve a more detailed and phenomenological understanding of the hydrolysis process, more substrate and enzyme properties have to be considered in the kinetic models. While models which do so would be more robust, they would require more experimental data for validation due to the increase in the number of variables and parameters.

Product inhibition of cellulases (by cellobiose) is a phenomenon that can be quantified by independent experiments and can be alleviated with an excess of β-glucosidase (Bommarius et al., 2008). The overall structure of the kinetic models of enzymatic hydrolysis of cellulose and lignocellulose is not affected by the inclusion of

product inhibition parameters. The phenomenon has been previously reviewed in 2002 (Lynd et al., 2002) and 2004 (Zhang and Lynd, 2004), and the state of the art in modeling product inhibition has not advanced since then. Therefore, in this chapter, the various expressions used for product inhibition are not discussed.

2.2 Model classes and classification

As shown in Figure 1 (Chapter 1), cellulose hydrolysis by cellulases is heterogeneous in nature, making classical enzyme kinetics models an oversimplification. Based on the fundamental approach and methodology used, the models can broadly be divided into four classes: empirical models (2.2.1), Michaelis-Menten based models (2.2.2), models accounting for adsorption (2.2.3), and those models developed for soluble substrates (2.2.4.) (Table 1). In addition, there are a few models in the literature based on jamming and fractal kinetics (discussed in section 2.3.5).

2.2.1 Empirical models

Though empirical models are not applicable outside the conditions under which they are developed and do not provide any insight into the mechanistic details of the process, they help in quantifying the effects of various substrate and enzyme properties on hydrolysis. Table 1A provides a list of empirical models in the literature, along with their predicted and independent variables. These empirical models have been generally used to correlate hydrolysis with either the structural properties of the substrate or with time. Empirical models can be helpful in numerous ways:

a) They can help in understanding the interactions between the substrate properties. It has been shown that the effects of an individual substrate property such as crystallinity, lignin content, or acetyl content can depend on the levels of the other two (Chang and Holtzapple, 2000; Kim and Holtzapple, 2006; O'Dwyer et al., 2008).

b) Empirical models can be useful for initial rate estimations, which are important for resuspension experiments (described in Section 2.3.3) and Lineweaver-Burk plots (Lineweaver and Burk, 1934) used in the Michaelis-Menten models. The rate of hydrolysis decreases continuously over time and to extrapolate the rate to time zero, an empirical formulation is needed. This can be illustrated by the empirical expression developed by Ohmine et al. (1983), where the following equation was found to hold for Avicel (partially acid hydrolyzed microcrystalline cellulose) and tissue paper hydrolysis by the cellulase system from *Trichoderma viride*:

$$P = \left(\frac{S_o}{k}\right)\ln(1 + v_o kt/S_o) \quad (1)$$

where P is the product concentration, S_o is the initial substrate concentration, v_o is the initial rate, k is the rate retardation constant and t is time.

For enzymatic hydrolysis of cellulose, to avoid the effects of product inhibition at product concentrations equal to zero, initial rates are plotted on the y axis vs. the reciprocal of the substrate concentration (in the Lineweaver-Burk plot) (Beltrame et al., 1984; Gusakov et al., 1985; Huang, 1975; Maguire, 1977; Shen and Agblevor, 2008a). These initial rates can be estimated using empirical expressions, such as:

i) Differentiating expressions by Sattler et al. (1989) (equation 2) and Koullas et al. (1992) (equation 4) with respect to time to get equations 3 and 5:
Sattler et al. (1989):

$$Y = (C_a + C_b) - C_a e^{-k_a t} - C_b e^{-k_b t} \quad (2)$$

$$\left.\frac{dY}{dt}\right|_{t=0} = C_a k_a + C_b k_b \quad (3)$$

where Y is the concentration of hydrolyzed cellulose, C_a and C_b are concentrations of easily and difficult hydrolysable parts of cellulose respectively, k_a and k_b are the rate constants of the first order hydrolysis of easily and difficult hydrolysable parts of cellulose, t is time, dY/dt (t=0) is the initial rate.

Koullas et al. (1992):

$$x = x_{max} \frac{t}{t_{1/2} + t} \quad (4)$$

$$\left.\frac{dx}{dt}\right|_{t=0} = \frac{x_{max}}{t_{1/2}} \quad (5)$$

where x is the conversion of cellulose to glucose, x_{max} is the maximum conversion, $t_{1/2}$ is the time required for 50% conversion, t is time, dx/dt (t=0) is the initial rate.

ii) Estimating v_o in the expression by Ohmine et al. (1983) (see above, equation 1).

c) When large data sets are available, statistical models can be used to optimize reaction conditions (Kim et al., 2008; Vásquez et al., 2007). Two examples employed response surface methodology to find optimal levels (to maximize cellulose conversion to glucose) of the operating conditions (pH, temperature, enzyme loading and solid percentage by Vásquez et al. (2007), pH, temperature and enzyme concentration by Kim et al. (2008)). Response surface methodology has also been used for optimizing cellulase mixtures (Berlin et al., 2007; Zhou et al., 2009a). Using steam-exploded corn stover as the substrate, Zhou et al. (2009a) optimized the composition of a mixture of *T. viride* cellulases (Cel7A, Cel6A, Cel6B, Cel7B, Cel12A, Cel61A and β-glucosidase) to maximize glucose production. O'Dwyer et al. (2008) developed a neural network model

to predict conversion levels as a function of crystallinity index, lignin content and acetyl content using data from 147 poplar wood samples. Such models which interpolate over a large range of the predicted and independent variables can be considered to have robust parameter values and can be useful for designing processes under various conditions.

2.2.2 Michaelis-Menten based models

The Michaelis-Menten scheme (Michaelis and Menten, 1913) is based on mass action laws that hold for homogenous reaction conditions and hence cannot be directly applied to the heterogeneous reaction conditions of enzymatic hydrolysis of insoluble cellulosic substrates. The excess substrate to enzyme ratio condition ([S] >> [E]), which is usually employed for the quasi-steady state assumption (Laidler, 1955; Schnell, 2003) is not achieved since the fraction of cellulose accessible for adsorption ranges from 0.002 – 0.04 (Hong et al., 2007). The excess substrate condition, even if ever achieved initially, could not be retained at higher conversions as the substrate gets depleted. It has also been pointed out by Lynd et al. (2002) that the concentration of adsorbed cellulase depends on the substrate concentration and that dual saturation is possible by keeping the enzyme or substrate concentration high; these features are not characteristic of Michaelis-Menten kinetics. Cellulose hydrolysis is a heterogeneous reaction occurring on the substrate surface and is therefore a reaction occurring in dimensions less than three. For heterogeneous reaction systems, classical chemical kinetics assumption of uniformly mixed systems does not hold, resulting in apparent rate orders, time-dependent rate constants, and non-uniform concentration variation of reacting species in the fractal or dimensionally restricted media (Anacker and Kopelman, 1987; Kopelman, 1988; Kopelman, 1986). Such a behavior is termed fractal kinetics. Monte Carlo simulations have corroborated that the quasi-steady state assumption cannot be applied in these reaction systems (Berry, 2002). Conversion of cellobiose to glucose by β-glucosidase,

however, can be modeled by Michaelis-Menten kinetics since it is a homogeneous reaction.

However, Michaelis-Menten models in the literature fit the experimental data very well under the conditions they were developed. Bezerra and Dias (2004) have tested eight different Michaelis-Menten models against data of Avicel hydrolysis by *T. reesei* Cel7A for 24 different substrate-to-enzyme ratios. A model with competitive inhibition by cellobiose was found to fit the data best. Reasons for the decreasing rates such as nonproductive cellulase binding, parabolic inhibition, and enzyme deactivation were shown to be insignificant in comparison to substrate depletion and competitive inhibition. Another work on Avicel with a fungal cellulase system from *T. viride* (Ohmine et al., 1983), however, had shown earlier that the same Michaelis-Menten model, incorporated with changes due to crystallinity and enzyme deactivation too, over-predicted the hydrolysis data. It was therefore suggested that either the kinetic scheme of the reaction is completely different or rate-retarding factors related to substrate heterogeneity are involved. The substrate heterogeneity factors are analyzed in section 2.3 ('Rate limitations and decreasing rates with increasing conversion').

2.2.3 Adsorption in cellulose hydrolysis models

Incorporation of adsorbed cellulase concentration into hydrolysis models has been achieved mainly in two ways: with the Langmuir adsorption isotherm, or with the help of kinetic equations. Fan and Lee (1983) observed constant amount of adsorbed cellulase per weight of cellulose along the hydrolysis and so a constant specific adsorption amount was used in their analysis. Movagarnejad et al. (2000) modeled the available number of active sites on the substrate surface as proportional to the surface area of the cellulose particles.

An example of a model employing the Langmuir adsorption isotherm is the one by Kadam et al. (2004). The adsorbed amount is given by:

$$E_b = \frac{E_{max} K_{ad} E_f S}{1 + K_{ad} E_f} \tag{6}$$

where E_b is the bound enzyme concentration, E_f is the free enzyme concentration, K_{ad} is the dissociation constant for adsorption, S is the substrate concentration, and E_{max} is the maximum adsorption capacity in amount of cellulase per amount of cellulose.

An example of the models using kinetic equations for the amount of enzyme adsorbed is the one by Gan et al. (2003) where the following equations were used for the adsorbed species:

$$E + S_c \underset{k_{sc2}}{\overset{k_{sc1}}{\rightleftharpoons}} E^*S_c \tag{7}$$

$$\frac{dC_{E^*S_c}}{dt} = k_{sc1} C_E C_{S_c} - k_{sc2} C_{E^*S_c} - k_p C_{E^*S_c} \tag{8}$$

where E is the enzyme, S_c is the active cellulose, E^*S_c is the enzyme-cellulose complex, C_E is the enzyme concentration, C_{E^*Sc} is the enzyme-cellulose complex concentration, C_{Sc} is the active cellulose concentration, k_{sc1} is the adsorption constant on active cellulose, k_{sc2} is the desorption constant on active cellulose, and k_p is the product formation constant.

Some of the models (Al-Zuhair, 2008; Brown and Holtzapple, 1990; Converse et al., 1988; Drissen et al., 2007; Fan and Lee, 1983; Gan et al., 2003; Huang, 1975; Kadam et al., 2004; Lin et al., 2005; Moon et al., 2001; Nidetzky and Steiner, 1993; Peri et al., 2007; Shen and Agblevor, 2008a; South et al., 1995; Wald et al., 1984) assume

instantaneous substrate-enzyme complex formation (fully productive adsorption), so the adsorbed amount of cellulase is the same as the amount of substrate-enzyme complexes. Some others (Asenjo, 1984; Converse and Optekar, 1993; Ding and Xu, 2004; Holtzapple et al., 1984; Liao et al., 2008; Luo et al., 1997; Ryu et al., 1982) assume an additional kinetic step on the substrate surface after cellulase adsorption, as did Luo et al. (1997), where the adsorbed cellulase combines with substrate to form a cellulase-substrate complex:

$$E_c' + C \xleftrightarrow{K_1} E_c'C \tag{9}$$

where E_c' is the adsorbed enzyme on the active sites, C is cellulose, , K_1 is the equilibrium constant, and $E_c'C$ is the cellulase-substrate complex. Brown and Holtzapple (1990) and Holtzapple et al. (1984) used the quasi-steady state assumption for the adsorbed enzyme and the substrate-enzyme complex species.

While isotherms other than the Langmuir isotherm, such as the Langmuir-Freundlich isotherm (Medve et al., 1997) and two-site models (Medve et al., 1998; Medve et al., 1997; Ståhlberg et al., 1991), have been shown to fit the data, only the Langmuir isotherm has been used in hydrolysis models. However, the Langmuir isotherm should only be used as a mathematical expression since its underlying assumptions (reversibility, non-interacting adsorbed species, homogenous binding sites and uniform composition of adsorbed cellulase mixture) may not be valid in all situations (Zhang and Lynd, 2004).

While using the Langmuir isotherm or any other mathematical expression for calculating the adsorbed amount of enzyme during hydrolysis, an implicit assumption is that the adsorption equilibrium is established very fast as compared to the hydrolysis step. According to Steiner et al. (1988), this assumption may not be valid under all

experimental conditions. The time to reach equilibrium adsorption has been estimated to be of the order of 5-60 minutes (Bader et al., 1992; Beldman et al., 1987; Ghose and Bisaria, 1979; Kim et al., 1994; Medve et al., 1998; Medve et al., 1994; Nidetzky et al., 1994a; Ståhlberg et al., 1991; Steiner et al., 1988). Though the time required for complete hydrolysis of cellulose (100% conversion) is usually 25-100 hours (Bertran and Dale, 1985; Bommarius et al., 2008; Gregg and Saddler, 1996; Nutor and Converse, 1991; Tu et al., 2007), the time for low conversion levels is two to three orders of magnitude lower (Bommarius et al., 2008; Hong et al., 2007; Nutor and Converse, 1991; Väljamäe et al., 1998). Also, use of the same isotherm at all time points of the reactions assumes that adsorption characteristics of the substrate-enzyme system do not change. If both assumptions (equilibrium of the adsorption and a single isotherm valid for all conversion levels) hold true, then the amount of enzyme adsorbed per unit weight of the substrate can only increase (see below).

Mass balance on the enzyme gives:

$$S*E_{ads} + E_f = E_{tot} \quad (10)$$

where S is the substrate concentration (g/L or equivalent units), E_{ads} is the specific adsorption amount (g cellulase/g cellulose or equivalent units), E_f is the free enzyme concentration (g/L or equivalent units), and E_{tot} is the total enzyme concentration (g/L or equivalent units).

Therefore, it follows that $E_{ads} = (E_{tot} - E_f)/S \quad (11)$

E_{ads} is also given by the adsorption isotherm:

$$E_{ads} = E_{max}K_{ad}E_f/(1+K_{ad}E_f) \quad (12)$$

Thus E_{ads} is determined by the intersection of equations (11) and (12). As S decreases along the course of hydrolysis, the magnitude of the slope increases and E_{ads} increases (Figure 3).

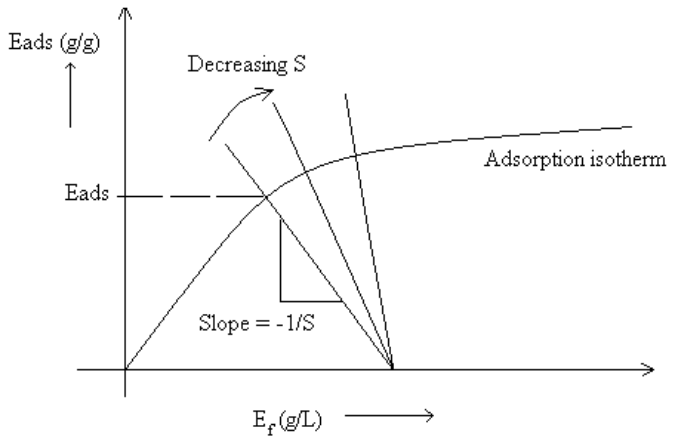


Figure 3. Determination of E_{ads} from the intersection of the mass balance equation and the Langmuir isotherm.

However, it is seen that E_{ads} does not monotonically increase with conversion, for both pure cellulosic substrates (Fan and Lee, 1983; Huang, 1975; Jeoh et al., 2006; Kurakake et al., 1995; Nidetzky and Steiner, 1993; Steiner et al., 1988) and lignocellulosic substrates (Kurakake et al., 1995; Liao et al., 2008; Nutor and Converse, 1991; Shen and Agblevor, 2008a; Steiner et al., 1988).

Hong et al. (2007), working with Avicel, have shown that the maximum adsorbable amount (E_{max} in the Langmuir isotherm) decreases with conversion. Empirical equations have also been developed for the changing concentration of adsorbed enzyme during the hydrolysis reaction (Kurakake et al., 1995). Lignin and hemicellulose act as barriers to cellulases to reach the cellulose core, and thus the changes in adsorption characteristics will be more pronounced for lignocellulosic substrates as compared to pure cellulosic substrates. The adsorption characteristics can depend on the type of substrate used, and since the isotherm parameters can change with conversion, it is important to validate the model against a measured amount of adsorbed cellulase during the hydrolysis. Shao et al. (2009a), Liao et al. (2008), and Nidetzky and Steiner (1993) incorporate adsorption and validate their models against experimental values for adsorbed

cellulases during hydrolysis. Using paper sludge as the substrate, Shao et al. (2009a) modeled adsorption of cellulases by the rate equations 12 and 13, and found that the same adsorption parameters fitted the data till 65% conversion; whereas, Liao et al. (2008), who used lignocellulosic material from dairy manure as the substrate, represented the change in the adsorption constant by an empirical expression (in time) fitted to the experimental data of adsorbed cellulase (equation 14).

$$r_{CE} = k_{fc}[E_f](1+\sigma_C)[C_f] - \frac{k_{fc}}{K_C}[CE] \quad (12)$$

$$r_{LE} = k_{fl}[E_f](1+\sigma_L)[L_f] - \frac{k_{fl}}{K_L}[LE] \quad (13)$$

where CE denotes cellulose enzyme complex, LE denotes lignin enzyme complex, r_{CE} and r_{LE} denote the rate of formation of cellulose enzyme complex and lignin enzyme complex respectively, σ_C and σ_L denote the adsorption capacities of cellulose and lignin respectively, k_{fc} and k_{fl} are the dynamic adsorption constants, $[E_f]$, $[C_f]$ and $[L_f]$ are concentrations of free enzyme, cellulose and lignin respectively, K_C and K_L are the adsorption constants.

$$K = \frac{at}{b+t} \quad (14)$$

where a and b are empirical constants, t is time, and K is the adsorption constant.

Nidetzky and Steiner (1993), who used four different cellulosic substrates (Sigmacell, Avicel, alpha-cellulose, cotton liners), represented the adsorption-desorption process over the conversion range as three phases: phase 1 where cellulases are adsorbed rapidly, phase 2 where desorption is linearly proportional to substrate conversion, and

phase 3 where desorption occurs at a very low rate. The three works mentioned here used different substrates and the validation of the adsorption model was done independent of the kinetic model, so that the differences in the adsorption model fitting cannot be attributed to the different natures of the overall kinetic models. Cellulases were the only enzymes used in these works, so the differences in adsorption characteristics cannot be expected to be due to enzymes but are mainly due to the different nature of the substrates. The adsorption characteristics can thus be substrate-dependent.

2.2.4 Models on soluble cello-oligosaccharides

Only a few models have been published on the cellulase hydrolysis of soluble cello-oligosaccharides (Table 1 C). While these models can be used to describe the hydrolysis of soluble substrates, extension to insoluble substrates is not straightforward. This is mainly because of the heterogeneous nature of cellulase action on insoluble cellulosic substrates. The concentration and distribution of accessible chain ends in insoluble substrates is also not known, and is probably subject to improvements in measurement techniques that can detect chain ends and their parts accessible/exposed to enzymes. This is a critical property, especially when it comes to modeling cellulose hydrolysis as polymer degradation (Okazaki and Moo-Young, 1978). With a recent study claiming that cellulose hydrolysis leads to the production of cello-oligosaccharides that are possibly not degraded by endoglucanases and exoglucanases (Gupta and Lee, 2008), models on soluble cellulosic substrates might provide more insight into the hydrolysis mechanism.

Recently, Ting et al. (2009) developed a stochastic model which gave insights into the modularity of the cellulases. The catalytic domain (CD) and the cellulose binding domain (CBD) were modeled as random walkers whose dynamics were coupled by the compression/expansion of the linker and lifting of cellulose chain from the substrate surface. For simplicity, only the major governing equation is shown:

$$\frac{dP(x,r,t)}{dt} = k_C(r+1)P(x-1,r+1,t)+k_{B-}(r+1)P(x,r+1,t) \\ +k_{B+}(r-1)P(x,r-1,t)-[k_C(r)+k_{B+}(r)+k_{B-}(r)]P(x,r,t) \quad (15)$$

where x denotes the position of CD, r is the separation between the CD and CBD, P denotes the probability of CD being at position x (the first entry in the parenthesis) with separation r (second entry in the parenthesis) from the CBD at time t (third entry in the parenthesis), $k_C(r)$ is the transition probability per unit time (for the CD) to move towards the CBD to a distance of r-1 from r, $k_{B+}(r)$ is the probability of the CBD to move away from the CD to a distance of r+1 from r, $k_{B-}(r)$ is the probability of the CBD to move towards the CD to a distance of r-1 from r.

The constants in the equations are then described by the energy dynamics arising from the compression/expansion of the linker, energy dynamics of hydrolysis, and chain disruption from the crystalline substrate surface. It was found that the linker flexibility/stiffness was an important factor governing the hydrolysis rates, as was the intrinsic hydrolytic activity of the CD. This is the first kinetic model which has attempted to explain the dynamics of the cellulose hydrolysis process by capturing the modular nature of the cellulases.

Table 1 (A) Empirical models (BG – β-glucosidase)

Reference	Y (Predicted Variable)	X (Independent Variable)	Substrate	Enzyme source	Validation range of conversion
Gharpuray et al. (1983)	Extent of Hydrolysis (after 8 hours)	Crystallinity, Lignin, specific surface area	Pretreated winter crop wheat straw	*T. reesei*	<70%
Ohmine et al. (1983)	Conversion	Time	Avicel	*T. viride*	>70%
Sattler et al. (1989)	Conversion	Time, **fractions of easily and difficult hydrolysable part**	Pretreated poplar wood	Celluclast + BG (Novo, Denmark)	>70%
Koullas et al. (1992)	Conversion, maximum conversion	Time, lignin, crystallinity	Ball milled Avicel, ball milled alkali-treated straw, ball milled wheat straw, alkali-treated wheat straw	*Fusarium oxysporum*	>70%
Ooshima et al. (1991)	Conversion, hydrolysis rate, adsorbed enzyme	Time	Avicel	*T. viride*	<70%
Kurakake et al. (1995)	Conversion, hydrolysis rate, adsorbed enzyme	Time	Avicel, pretreated Wilner hardwood	*T. reesei, T. viride*	>70%
Parajó et al. (1996)	Conversion	Time, **fractions of easily and difficult hydrolysable parts**	NaOH pretreated pine wood	*T. reesei* + BG	<70%
Tarantili et al. (1996)	Conversion	Time, maximum conversion, time for achieving half of maximum conversion	Ball milled Avicel, filter paper, Greek purified cotton and hot-alkali-delignified wheat straw	*Fusarium oxysporum* and *Neurospora crassa*	<70%
Moldes et al. (1999)	Maximum rate of cellulose conversion, max. rate of glucose generation	Enzyme to substrate ratio, liquor to solid ratio	Pretreated wood chips	Celluclast (Novo Denmark)	<70%
Chang and Holtzapple (2000)	1 hour and final conversions of glucan and xylan content	Lignin content, acetyl content, glucan content, crystallinity index	Hybrid poplar, bagasse and switchgrass	Cytolase (cellulase from Environmental BioTechnologies, Santa Rosa, CA) + BG	>70%
Park et al. (2002)	Conversion	Time, enzyme concentration	Waste office paper	*T. viride, Acremonium cellulolyticus*	>70%
Laureano-Perez et al. (2005)	Initial hydrolysis rate, 72 h extent of hydrolysis	Crystallinity, Spectroscopic features	Corn Stover	Cellulase from NREL + BG	>70%
Kim and Holtzapple (2006)	Hydrolysis yields of glucan, xylan and holocellulose	Residual lignin	Pretreated corn stover	Spezyme CP from NREL + BG	>70%
Vásquez et al. (2007)	Glucose concentration	pH, enzyme loading, temperature, solid percentage	Acid hydrolyzed sugarcane bagasse	GC 220 (Genencor International, Inc.)	<70%

Table 1(A) (continued)

Reference	Y (Predicted Variable)	X (Independent Variable)	Substrate	Enzyme source	Validation range of conversion
Berlin et al. (2007)	Glucan to glucose and xylan to xylose conversion	Weights of xylanse, pectinase and β-glucosidase	Milled corn stover, dilute acid pretreated corn stover	Celluclast 1.5L +BG (Novozymes), xylanse and pectinase (Genencor International)	>70%
O'Dwyer et al. (2008)	Slopes and intercepts of the graphs of 1h, 6, 72h glucan content vs enzyme loading	Crystallinity, Lignin and acetyl content	Pretreated poplar wood	*T. reesei*	>70%
Kim et al. (2008)	Reducing sugar concentration, ethanol concentration	pH, temperature, enzyme inoculation, reaction time	Food waste	Spirizyme Plus FG (Novozymes, Denmark)	-
Zhou et al. (2009a)	Glucose produced after 72 hours hydrolysis	Concentrations of Cel7A, Cel6A, Cel6B, Cel7B,Cel12A, Cel61A	Steam-exploded corn stover	*T. viride* (Cel7A, Cel6A, Cel6B, Cel7B, Cel12A, Cel61A) + BG	<70%

(In bold are shown the assumptions regarding the decrease in rates)

Table 1(B). Adsorption and Michaelis - Menten based models (M-M: Michaelis – Menten, PI – Product inhibition, QSS – Quasi Steady State assumption, Ads – Adsorption based approach, BG – β-glucosidase)

Reference	Methodology	Substrate	Enzyme source (purified component if any)	Declining rate reason (in addition to PI)	Conversion range for validation
Huang (1975)	Ads, QSS	Amorphous Solka Floc	*T. viride*	-	>70%
Suga et al. (1975)	M-M	Theoretical study			
Howell and Stuck (1975)	M-M	Solka Floc	*T. viride*	-	<70%
Maguire (1977)	Ads	Alpha cellulose fiber	*T. viride* Cellobiohydrolase (then known as the C_1 enzyme)	-	<70%
Howell (1978)	QSS	Solka Floc	*T. viride*	Enzyme inactivation	<70%
Okazaki and Moo-Young (1978)	QSS, M-M	Analytical Study			
Peiterson and Edward W. Ross (1979)	M-M, two phases – crystalline + amorphous	Ball milled delignified cellulose	*T. reesei* + BG	two phases – crystalline + amorphous	<70%
Ryu et al. (1982)	Ads, M-M	Solka Floc, Avicel, adsorbant cotton	*T. reesei*	Accessibility decreases with increase in CrI	<70%
Fan and Lee (1983)	Ads	Solka Floc	*T. reesei* + BG	Decrease in Substrate reactivity	>70%
(Asenjo, 1983; Asenjo, 1984)	Ads	Solka Floc	*T. viride*	Only a fraction is available for attack	<70%

Table 1(B) (continued)

Reference	Methodology	Substrate	Enzyme source (purified component if any)	Declining rate reason (in addition to PI)	Conversion range for validation
Beltrame et al. (1984)	M-M	Textile, cotton waste, pretreated pulp	*T. viride* + BG	-	<70%
Holtzapple et al. (1984)	Ads, QSS	Solka Floc	*T. viride* + BG	Accessibility is included as a parameter	<70%
Scheiding et al. (1984)	M-M	Avicel	*T. reesei* + BG	Enzyme deactivation, amorphous + crystalline fractions	<70%
Wald et al. (1984)	Ads, QSS, Apparent rate order	Rice straw	*T. reesei* + BG	-	<70%
Caminal et al. (1985)	M-M	Microcrystalline cellulose form Merk	Cellulase from Merk	Enzyme deactivation	>70% (only fitting)
Gusakov et al. (1985)	M-M	Chemically treated cotton stalks	*Trichoderma longibrachiatum* +BG	Enzyme inactivation, two phase substrate model	>70%
Converse et al. (1988)	Ads, QSS, rate constant time dependent	Microcrystalline cellulose form Merk	*T. viride*	Enzyme deactivation, bulk mass transfer limitation	>70% (only fitting)
Nakasaki et al. (1988)	M-M	Filter paper	Meicelase CEPB-5081	Some fraction non-degradable	~70%
Converse et al. (1990)	Ads, accessibility characterized by surface area (scientific note)	Pretreated wood	*T. reesei* + BG	Change in surface area of substrate	Fitted to initial hydrolysis rate
(Philippidis et al., 1993; Philippidis et al., 1992)	Ads	Alpha cellulose, cellobiose and gluconolactone	*T. reesei* + BG	Enzyme deactivation, adsorption of cellulase and β-glucosidase onto lignin, substrate reactivity coefficient included for substrate reactivity	<70%
Converse and Optekar (1993)	Ads	Avicel	data from Woodward et al. (1988b) (*T. reesei* (Cel6A, Cel7A, Cel5A))	-	<70%
Nidetzky and Steiner (1993)	Ads, M-M, two phase substrate	Sigmacell, Avicel, alpha-cellulose, cotton liners	Celluclast + BG (from Novo, Denmark)	Enzyme desorption, two phases of substrate	>70%
Nidetzky et al. (1993)	Psuedo 2nd order reaction wrt substrate	Wheat straw	Celluclast + BG (from Novo, Denmark)	-	>70%
Nidetzky et al. (1994b)	M-M	Whatman no. 1 Filter paper	*T. reesei* (Cel7A, Cel6A, Cel 7B) +BG	-	<70%
South et al. (1995)	Ads	Data of Nutor and Converse (1991)	*T. reesei*	Decrease in substrate reactivity	>70%
Luo et al. (1997)	Ads.	Pretreated corn cob	*Trichocherium reesei*/ *Aspergillus niger*	Enzyme deactivation	>70%
Fenske et al. (1999)	Monte Carlo simulations	Theoretical study			
Moldes et al. (1999)	M-M, Empirical	Pretreated wood chips	Celluclast (from Novo, Denmark)	-	<70%
Schell et al. (1999)	Same as Philippidis et al. (1992)	Dilute acid pretreated Douglas fir	Iogen super clean cellulase	Same as Philippidis et al., (1992)	>70%
Movagarneja d et al. (2000)	Ads, shrinking particle theory	Microcrystalline cellulose from Merck	Celluclast + BG (from Novo, Denmark)	Inactive complexes formed on substrate	<70%

Table 1(B) (continued)

Reference	Methodology	Substrate	Enzyme source (purified component if any)	Declining rate reason (in addition to PI)	Conversion range for validation
Moon et al. (2001)	Ads	Alpha cellulose, ball milled and untreated steam exploded wood	Celluclast + BG (from Novo, Denmark)	Substrate reactivity, enzyme deactivation	<70%
Pettersson et al. (2002)	Ads	Data from Stenberg et al. (2000) (Steam-pretreated softwood)	Cellulcast 2L + BG (from Novo, Denmark)	Decrease in cellulose specific surface area, adsorption of cellulase and β-glucosidase onto lignin	>70%
Gan et al. (2003)	Ads	Alpha-cellulose (Sigma C802)	*T. reesei*	Inert fraction of cellulose, enzyme deactivation	<70%
(Movagharn ejad, 2005; Movagharnej ad and Sohrabi, 2003)	Ads, shrinking particle theory	Cellulosic waste materials	Celluclast + BG (from Novo, Denmark)	Inaccessibility of active sites to enzyme	<70%
Bezerra and Dias (2004)	Evaluation of M-M models	Avicel	*T. reesei* (Cel7A)	-	<70%
Shen et al. (2004)	Ads	Dried cotton, viscose and flax yarns	*Trichoderma pseudokoningii*	-	<70%
Ding and Xu (2004)	Ads, QSS	PASC, Avicel, PCS	*T. reesei* (Cel7A and Cel7B) and *H insolens* (Cel6A and Cel7A)	-	Evaluation of accessibility with initial rates only
Kadam et al. (2004)	Ads	Pretreated corn stover	CPN commercial cellulase (Iogen Corp.) + BG	Substrate reactivity	<70%
Lin et al. (2005)	Ads	Cellulose powder 101-F (Sigma, USA)	*T. reesei* + BG	Adsorbed enzyme converted irreversibly into inactive complex	<70%
Kipper et al. (2005)	Active- site titration theory	Bacterial cellulose (BC), bacterial microcrystalline cellulose, endoglucanase pretreated BC, amorphous cellulose	*T. reesei* (Cel7A, Cel6A, Cel5A)	-	<70% (Note: the purpose of the study was to check for burst kinetics hence data for low hydrolysis times was used)
Shin et al. (2006)	M-M	Alpha cellulose, ball milled and untreated steam exploded wood	Data from Moon et al. (2001) (Celluclast + BG from Novo, Denamrk)	Inhibition by lignin, enzyme deactivation	<70%
Ljunggren (2005)	Ads	Pretreated spruce, pretreated sugar cane bagasse	Celluclast + BG (from Novo, Denmark)	Enzyme deactivation, β-glucosidase adsorption to lignin	>70%
Zhang and Lynd (2006)	Ads	PASC, Avicel, bacterial cellulose, cotton, filter paper	Results compared with literature on *T. reesei* cellulase system (Cel7A, Cel6B, Cel7B)	-	Parameter values taken from literature, initial rates compared with data from Wood (1975)

Table 1(B) (continued)

Reference	Methodology	Substrate	Enzyme source (purified component if any)	Declining rate reason (in addition to PI)	Conversion range for validation
Drissen et al. (2007)	M-M, Ads	Avicel, Whatman Filter paper, wheat straw	Cellubrix (Novozymes Corp., Denmark) + BG	Enzyme deactivation, decreasing reactivity of substrate	<70%
Peri et al. (2007)	Ads	Non crystalline cellulose (prepared from cotton and α-cellulose), α-cellulose	Spezyme CP (Genencor)	-	>70%
O'Dwyer et al. (2007)	Same model as Holtzapple et al. (1984)	Lime-pretreated corn stover	*T. reesei* + BG	-	>70%
Al-Zuhair (2008)	Ads	Highly crystalline wood shavings, carboxymethylcell -ulose (CMC)	*Aspergillus niger*	Two phase substrate	<70%
Liao et al. (2008)	Ads	Lignocellulosic material from dairy manure	Celluclast + BG (from Signma)	Decreasing substrate reactivity	>70%
Shen and Agblevor (2008a)	Ads, QSS	Steam-exploded cotton gin waste	Novozymes NS 50052 (from Novozymes) and Spezyme AO3117 (from Genencor International)	Enzyme deactivation	<70%
Shen and Agblevor (2008b)	Ads, QSS	Cotton gin waste, recycled paper sludge	Spezyme AO3117 (Genencor International)	Enzyme deactivation	<70%
(Shao et al., 2009a; Shao et al., 2009b)	Ads	Waste paper sludge	Spezyme CP (Genencor) + BG (Sigma-Aldrich)	Decreasing substrate reactivity	>70%
Zheng et al. (2009)	Ads	Creeping wild ryegrass	Celluclast + BG (Novozymes Inc.)	Decreasing substrate reactivity, adsorption to lignin	>70%
(Zhou et al., 2009b; Zhou et al., 2009c)	ODEs[a] for evolving substrate morphology	Theoretical study			
Levine et al. (2010)	Ads	Data of Medve et al. (1998). Avicel	CBHI, EG2	Decreasing accessibility	<70%
Zhang et al. (2010)	Ads, QSS	Steam exploded wheat straw	Crude cellulase powder (Shanghai Bio Life Science & Technology Co., Ltd.)	Enzyme deactivation	<70%
Praestgaard et al. (2011)	Ads	Amorphous cellulose generated from Sigmacell 20	*T. reesei* (Cel7A)	Enzymes getting stuck at obstacles	<70%
(Podkaminer et al., 2011)	Ads	Avicel PH-105	Spezyme CP (Genencor), BG (Novozyme 188)	Decreasing substrate reactivity, enzyme inactivation by ethanol	>70%

[a] Ordinary differential equation

Table 1(C) Models on soluble cello-oligosaccharides (DP - Degree of polymerization)

Reference	Substrates	Enzyme source (pure component if any)
Fujii and Shimizu (1986)	Carboxy methyl cellulose and hydroxyl ethyl cellulose	*Trichoderma Koningii*
Schmid and Wandrey (1989)	Cellodextrins with chain lengths of 2 (cellobiose) to 6 (cellohexaose)	β-glucosidase from *T. reesei*
Nassar et al. (1991)	Model validated with data from Schmid and Wandrey (1989)	
DeanIII and Rollings (1992)	Dextran (polysaccharide with α-1,6-glycosidic linkages)	Endodextranase (from a *Penicillium* species from Sigma) and exodextranase from *Arthrobacter globiformis*
Nidetzky et al. (1994c)	Cello-oligosaccharides with DP up to 8	*T. reesei* (Cel7A and Cel6A)
Harjunpää et al. (1996)	Cello-oligosaccharides with DP 4 – 6	*T. reesei* (Cel6A)

Table 1(D) Models on jamming and fractal kinetics

Reference	Substrate	Enzyme source (pure component if any)	Range of validation
Väljamäe et al. (2003)	Bacterial cellulose	*T. reesei* (Cel7A, Cel5A)	(<10%) (Note: The objective was to fit the data to find the parameter h, representing the fractal dimension)
Xu and Ding (2007)	Avicel and PASC	*H. insolens* (Cel7A), *T. reesei* (Cel7A)	>70%
Wang and Feng (2010)	Literature data (Bommarius et al., 2008; Borjesson et al., 2007; Kim and Lee, 2005; Wu and Ju, 1998)		>70%

2.3. Rate limitations and decreasing rates with increasing conversion

Even after alleviating product inhibition from cellobiose, cellulase activities and hydrolysis rates fall precipitously as the reaction proceeds (Bommarius et al., 2008). To be able to increase the rates, the various bottlenecks in cellulose hydrolysis need to be elucidated.

The contributing factors to decreasing rates (other than product inhibition) accounted for in the existing models include (see Table 1): a) enzyme deactivation (2.3.1), b) biphasic composition of cellulose (2.3.2), c) decrease in substrate reactivity (2.3.3), d) decrease in substrate accessibility (2.3.4), e) jamming and fractal kinetics (2.3.5), and f) decrease in the synergism between cellulases (2.3.6). For substrates containing lignin and other non-cellulosic components, additional factors such as

inaccessibility caused by lignin and adsorption of cellulases to lignin will contribute to rate limitations; these aspects are discussed in section 2.5.

2.3.1. Enzyme deactivation

While enzyme deactivation has often been modeled as a first order process with respect to the total enzyme concentration (Caminal et al., 1985; Drissen et al., 2007; Ljunggren, 2005; Luo et al., 1997; Moon et al., 2001; Oh et al., 2000; Philippidis et al., 1993; Philippidis et al., 1992; Schell et al., 1999; Shin et al., 2006) inactivation of the adsorbed enzyme only has also been considered (Converse et al., 1988; Gusakov et al., 1985; Howell, 1978; Lin et al., 2005; Scheiding et al., 1984). Gan et al. (2003) considered the loss of enzyme due to shear force. Shen and Agblevor (2008b), and Shen and Agblevor (2008a) assumed enzyme deactivation to be a second-order reaction.

As an example of the enzyme deactivation of the adsorbed enzyme, Converse et al. (1988) used the following reaction representing enzyme deactivation:

$$E_a \underset{k_2}{\overset{k_1}{\rightleftharpoons}} E_d \tag{16}$$

where E_a is the actively adsorbed enzyme, E_d is the inactively adsorbed enzyme, and k_1 is the inactivation rate constant, k_2 is the reactivation rate constant.

Enzyme deactivation has also been related to enzyme clogging in an erosion model (Väljamäe et al., 1998), where the cellobiohydrolases become stuck on the substrate surface when surrounding cellulose chains prevent further processive action. Jalak and Valjamae (2010) measured the observed catalytic rate constant (k_{obs}), by diving the instantaneous rate by the concentration of cellobiohydrolases whose active site was occupied by cellulose chains, $[CBH]_{OA}$ (OA – occupied active site). $[CBH]_{OA}$ was

observed to remain constant, whereas k_{obs} decreased rapidly in the initial stages. As some fraction of $[CBH]_{OA}$ can be unproductively bound, the decrease in k_{obs} was attributed to the increase in this fraction. According to the authors, in the initial 'burst phase', the cellobiohydrolases make a quick run hydrolyzing cellulose until they encounter an obstacle, resulting in unproductive binding. Based on the observation that the observed processivities of *T. reesei* Cel7A and *Phanerochaete chrysosporium* Cel7D were an order of magnitude lower than their intrinsic processivities, Kurasin and Valjamae (2011) concluded the number of catalytic cycles performed before desorption is limited by the obstacle free length on the cellulose chain. Recently, Praestgaard et al. (2011) modeled this phenomenon, and when fit to data with amorphous cellulose, could only explain the hydrolysis data in the initial stages. This implies that either there are more phenomena responsible for the rate decline, or that state of the art of modeling cellulose hydrolysis needs to be improved.

Through restart hydrolysis experiments, Yang et al. (2006) also suggested stopping or slowdown of the enzymes on the substrate surface to account for the reaction rate slowdown. Eriksson et al. (2002) showed that thermal enzyme instability and product inhibition are not the major causes for the reduction in rates. The authors proposed a model where cellobiohydrolases encounter obstacles during their processive action while endoglucanases partially remove this hindrance by hydrolyzing the responsible cellulose chains. This study however, was performed with steam-pretreated spruce, a lignocellulosic substrate where non-cellulosic parts can also possibly act as obstacles to enzymes.

2.3.2. Two-phase substrate

Under the assumption of a two-phase substrate, the more reactive part reacts faster resulting in a decrease in its overall fraction and a concomitant decrease in the overall reaction rate with time. Some works suggested that the amorphous part of cellulose reacts

first (accompanied by an increase in crystallinity) (Chen et al., 2007; Gan et al., 2003; Lee and Fan, 1983; Mansfield and Meder, 2003; Medve et al., 1994; Ohmine et al., 1983; Ooshima et al., 1983; Szijártó et al., 2008; Väljamäe et al., 1999; Zhang et al., 1999), while constant (Cateto et al., 2011; Hall et al., 2010a; Lenz et al., 1990; Puls and Wood, 1991) and decreasing crystallinity (Mansfield and Meder, 2003) along conversion have also been reported. Zhang and Lynd (2004), and Mansfield et al. (1999) reported this dichotomy as well. Hall et al. (2010a) showed that although cellulose crystallinity does not change with conversion, it is a key predictor of hydrolysis rates, and there exists a strong linear relationship between the initial rates and degree of crystallinity. One source of dissonance in the literature is the lack of techniques for consistent measurement of cellulose crystallinity, as highlighted in Park et al. (2010). With improvements in measurement techniques and calculation methods for cellulose crystallinity (Bansal et al., 2010; Barnette et al., 2011; Park et al., 2009), the relationship between cellulose crystallinity and hydrolysis rates is expected to become clearer.

Models assuming cellulose to be divided into crystalline and amorphous fractions have been proposed (Gusakov et al., 1985; Peiterson and Edward W. Ross, 1979; Ryu et al., 1982; Scheiding et al., 1984). These works, however, did not verify their assumptions by measuring the crystallinity of the substrate along conversion. Based on a Michaelis-Menten scheme of the biohydrolysis of amorphous and crystalline fractions, Ryu et al. (1982) obtained the following two equations:

$$\frac{v_{max}''}{K_M''} = \left(\frac{v_{max,c}}{K_{M,c}} - \frac{v_{max,a}}{K_{M,a}} \right) \Phi + \frac{v_{max,a}}{K_{M,a}} \tag{17}$$

$$\frac{1}{K_M''} = \left(\frac{1}{K_{M,c}} - \frac{1}{K_{M,a}} \right) \Phi + \frac{1}{K_{M,a}} \tag{18}$$

where v_{max}'' is the maximum apparent rate, $v_{max,c}$ is the maximum rate for crystalline fraction, $v_{max,a}$ is the maximum rate for amorphous fraction, K_M'' is the apparent Michaelis constant, $K_{M,c}$ is the Michaelis constant for crystalline fraction, $K_{M,a}$ is the Michaelis constant for amorphous fraction, and Φ is the fraction of crystalline phase.

The two-phase hypothesis, however, was emphasized to be a simplification of the true physical complexity of cellulose. Cellulose crystallinity was shown to affect the digestibility of cellulose by impacting its accessibility (Jeoh et al., 2007). In the same work, the specific activity of the adsorbed *T. reesei* Cel7A was higher on PASC (phosphoric acid swollen cellulose, amorphous cellulose) than on Avicel, implying either higher susceptibility of lesser crystalline cellulose towards hydrolysis or lesser non-productive adsorption. Crystallinity, therefore, is not an independent substrate property and can affect accessibility and reactivity of the cellulose sample.

It has also been assumed that a part of the substrate is inert, with the fraction of inert part remaining constant during conversion (Al-Zuhair (2008) – using CMC and wood shavings, Gan et al. (2003) – using cellulose). This fraction, however, was an assumed constant in the model equations. Models assuming a non-degradable fraction of cellulose have also been developed (Asenjo, 1983; Asenjo, 1984; Nakasaki et al., 1988). Based on the observation that 30% of the filter paper powder remained unreacted at long residence times (approximately 340 hours), Nakasaki et al. (1988) assumed the non-degradable fraction to be 0.3. Asenjo (1983), and Asenjo (1984), however, assuming the non degradable fraction to be 35% for Solka-Floc (a pure cellulosic substrate), did not validate the assumption of a non-degradable fraction by fitting the model predictions to experimental data up to the maximum theoretical conversion achievable (65%).
An empirical model by Parajó et al. (1996) took into account two parts of cellulose having different susceptibility towards enzymatic attack. According to Nidetzky and Steiner (1993), the presence of a) two parts of cellulose differing in their reactivity and b) a fraction of substrate that is non-degradable, are important factors affecting cellulose

enzymatic hydrolysis. Resuspension experiments (where enzymes are washed off the surface of the unreacted cellulose and the partially hydrolyzed substrate is subjected to cellulase hydrolysis under initial conditions) were used to show the existence of two fractions and the authors concluded that, though no physical property variation can explain the presence of two fractions, the possibility cannot be ruled out. Biphasic kinetics, however, seems unlikely to be the only cause for the rate slowdown.

2.3.3. Substrate reactivity

The change in substrate reactivity has been included in a number of models to explain the reduced digestibility of hydrolyzed cellulose, for both lignocellulosic and pure cellulosic substrates (Table 1B). Lee and Fan (1983) (pure cellulosic substrate) and Moon et al. (2001) (both pure cellulosic and lignocellulosic substrates) employed the initial hydrolysis rates from resuspension experiments of spent substrate to correlate 'relative digestibility' with conversion. As an example, Lee and Fan (1983) developed the following expression:

$$\phi = 1 - X^n \tag{19}$$

where ϕ is relative digestibility, X is conversion, and n is a parameter fitted with the help of resuspension experiments.

South et al. (1995) also expressed the reaction rate constant in terms of conversion:

$$k(x) = k(1-x)^n + c \tag{20}$$

where k is the reaction rate constant for hydrolysis, x is conversion, k(x) is the reaction rate constant at conversion x, n is the exponent of declining rate constant, c is a constant. n and c were estimated by approximating k(x) by the ratio of rate/adsorbed enzyme and fitting it with equation to conversion (x). This expression was later used in modeling SSF with staged reactors and intermediate feeding of enzyme and substrate (Shao et al., 2009a; Shao et al., 2009b). Based on the observation that the initial rates (for pretreated corn stover) followed a linear trend with the substrate concentration, Kadam et al. (2004) fitted the following equation for substrate reactivity:

$$R_s = \frac{S}{S_0} \tag{21}$$

where R_s is substrate reactivity, S is substrate concentration, S_0 is initial substrate concentration.

Liao et al. (2008) also used a similar expression (equation 22), but the parameters were not determined by independent experiments, and the reason for the use of this expression was stated to be for available cellulose for enzymes:

$$[C]_{eff} = \left(\frac{[C]}{[C]_0}\right)^{\lambda} [C] \tag{22}$$

where $[C]_{eff}$ is the concentration of cellulose available to enzymes, [C] is cellulose concentration, $[C]_0$ is initial cellulose concentration, λ is a constant. $([C]/[C]_0)^{\lambda}$ is equivalent to R_s in equation 21.

Although the inclusion of the rate constant or substrate reactivity as a function of conversion may fit the data well, a physical interpretation of the constants in these equations is not possible. The continuous decline in reactivity has been alternatively

explained by the consumption of a more reactive fraction of the substrate (Hong et al., 2007), leading back to the assumption of a biphasic substrate.

Various studies have used resuspension experiments to study the reactivity of partially converted cellulose (Desai and Converse, 1997; Drissen et al., 2007; Gusakov et al., 1985; Hong et al., 2007; Kumar and Wyman, 2009; Lee and Fan, 1983; Ooshima et al., 1991; Väljamäe et al., 1998; Yang et al., 2006; Zhang et al., 1999). As pointed out by Lynd et al. (2002) in 2002, there was no consensus regarding the decline of reactivity as observed in these experiments. Post 2002, through resuspension experiments, Hong et al. (2007) and Drissen et al. (2007) observed a decline in reactivity whereas Yang et al. (2006) did not. Generalization from the above results becomes more difficult since the enzyme system and substrate used were different. Kumar and Wyman (2009) showed that trends in reactivity and accessibility based on restart experiments can vary between lignocellulosic substrates depending on the pretreatment method.

2.3.4 Substrate accessibility

Due to the insoluble nature of cellulose, large domains are not exposed to cellulases in the reaction mixture during the hydrolysis reaction. Accessibility of cellulose can be characterized on the basis of adsorption. Cellulases can adsorb only to the accessible portion of the substrate, and this fraction is calculated based on the maximum adsorption capacity of the substrate (Hong et al., 2007; Zhang and Lynd, 2004):

$$F_a = 2\alpha A_{max} MW_{anhydroglucose} \tag{23}$$

where F_a is the fraction of the β-glycosidic bonds accessible to cellulase, α is the number of cellobiose lattice occupied by the cellulase, A_{max} is the maximum adsorption concentration of cellulase, and $MW_{anhydroglucose}$ is the molecular weight of anhydroglucose.

This fraction fell by approximately 50% from 0.002 until a conversion of around 85% (conversion of Avicel with *T. reesei* cellulase system) (Hong et al., 2007). In light of these findings, it might be important to take into account the reduced accessibility and adsorption capacity of the substrate as the conversion proceeds (also discussed in Section 2.2). Ding and Xu (2004) have estimated the 'kinetic accessibility' of Avicel and PASC to *T. reesei* and *H. insolens* cellulases from initial rate data (equations 24 and 25).

$$\varphi = \frac{[S]_0}{[S]_t} \tag{24}$$

where $[S]_0$ is the concentration of cellulose available to cellulases for productive adsorption, $[S]_t$ is the total concentration of cellulose, φ is the ratio of $[S]_0$ to $[S]_t$ and represents the kinetic accessibility of cellulose.

φ was estimated by the following expression:

$$\varphi = \beta \frac{[E]_0^{\,s}}{[S]_t} \tag{25}$$

$[E]_0$ denotes initial substrate concentration. At low $[E]_0$, v_0 is directly proportional to $[E]_0$ (i.e. $v_0=k[E]_0$)and at high concentrations v_0 is constant. The intersection of $v_0=k[E]_0$ and v_0 = constant gives $[E]_0^s$. β (=39) the number of cellobiosyl units covered by an adsorbed CBH.

The results showed that φ can be different for different cellulases for the same substrate, e.g. for Avicel, φ was 0.014 for Cel7A but only 0.0012 for Cel7B. The order of magnitude of φ and F_a is the same: F_a = 0.002 and φ = 0.0012-0.014 for four different enzymes.

The importance of productive adsorption can be illustrated by a simple analysis of the data from Zhang and Lynd (2005), and Hong et al. (2007):

Accessible fraction of the β-glycosidic bonds in Whatman Filter paper (as calculated by equation 27) ~ 0.0095, DP ~ 2000. Therefore:

$$[C]_r = \frac{1}{2000}[C]_b = 0.0005 * [C]_b \tag{26}$$

$$\text{and } [C]_a = 0.0095 * [C]_b \tag{27}$$

where $[C]_r$ is the concentration of reducing ends, $[C]_b$ is the concentration of all β-glycosidic bonds, $[C]_a$ is the concentration of accessible β-glycosidic bonds.
If all the chain ends are occupied at maximum adsorption, there would still be a large fraction of non-productively bound cellobiohydrolase given by:

$$\frac{[C]_a - [C]_r}{[C]_a} = \frac{0.0095 - 0.005}{0.0095} \sim 0.95 \tag{28}$$

As cellulose chains are hydrolyzed, the chains located below, which were not exposed to cellulases, can undergo hydrolysis. Based on this idea, accessibility parameters have been included in the rate equations (Al-Zuhair, 2008; Converse and Optekar, 1993; Gan et al., 2003; Wood, 1975). It is not clear whether it is possible to classify a part of the substrate in just two categories: accessible and inaccessible. Accessibility as a substrate property could possibly be a continuous variable.

2.3.5. Role of fractal kinetics in cellulase kinetics

Fractal kinetics is said to occur when reactions take place in spatially constrained media; such reaction conditions give rise to non-uniformly mixed reaction species, apparent rate orders, and time-dependent rate constants (Anacker and Kopelman, 1987; Kopelman, 1988; Kopelman, 1986). Since cellulase hydrolysis of insoluble cellulosic substrates can be thought of as a one-dimensional heterogeneous reaction along a cellulosic fiber, it can result in fractal kinetics. Though reactions occurring on a supported catalyst can be modeled using Langmuir-Hinshelwood kinetics (Fogler, 2005), fractal kinetics must be considered for catalytic reactions involving diffusion of two species (for bimolecular reactions) on the non-ideal substrate surfaces (surfaces with obstacles resulting in segregation of species, non-uniform concentrations).

Michaelis-Menten kinetics in fractal media was first studied using the power law formalism (Savageau, 1995), where the classical enzyme catalysis reaction (equation 29) in fractal media was described by apparent rate orders (equations 32 and 33).

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P \tag{29}$$

where E is enzyme, S is substrate, ES is enzyme-substrate complex, P is product, k_1 is the forward rate constant for the association of the enzyme and substrate, k_{-1} is the dissociation constant of the enzyme-substrate complex and k_2 is the product formation rate constant.

Classical equations –

$$\frac{d(\mathrm{ES})}{dt} = k_1 * \mathrm{E}*\mathrm{S} - (\mathrm{k}_{-1}+\mathrm{k}_2)(\mathrm{ES}) \tag{30}$$

$$\frac{d\mathrm{P}}{dt} = \mathrm{v}_\mathrm{p} = \mathrm{k}_2(\mathrm{ES}) \tag{31}$$

Power law equations –

$$\frac{d(\mathrm{ES})}{dt} = \alpha_1 * \mathrm{E}^{g_1} * \mathrm{S}^{g_2} - (\beta_1 + \alpha_2)(\mathrm{ES}) \tag{32}$$

$$\frac{d\mathrm{P}}{dt} = \mathrm{v_p} = \alpha_2(\mathrm{ES}) \tag{33}$$

where v_p is the product formation rate, α_1, $\alpha_2$and β_1 are new constants introduced for the power law formulation, g_1 and g_2 are the apparent rate order with respect to E and S.

Using Monte Carlo simulations, the classical enzyme kinetics scheme (equation 29), has been studied in two dimensions in the presence of surface obstacles by Berry (2002). The fractal nature of the reaction system was shown to increase as the obstacle density was increased. k_1 (rate constant of a bimolecular reaction requiring the diffusion of enzyme and substrate on the surface) was shown to decrease with time, whereas k_{-1} and k_2 were time-invariant, as the uni-molecular reaction did not require diffusion. It was also shown that the quasi-steady state assumption cannot be applied in these conditions. After adsorption, cellulases have to diffuse on the surface of the substrate to reach the reactive sites (a chain end in the case of cellobiohydrolases). The inaccessible and non-reactive portions of the substrate can be considered as obstacles increasing the fractal character of the hydrolysis reaction. The first work to study cellulose hydrolysis by fractal kinetics was performed by Väljamäe et al. (2003). Using an empirical first-order product formation equation for cellobiose production (equation 34), the parameter h, which represents the fractal dimension, was shown to increase with increasing substrate concentration for Cel7A core protein (catalytic domain only) plus Cel5A endoglucanase (0.1 to ~0.45) but to decrease for Cel7A intact protein plus Cel5A endoglucanase (0.6 to ~0.35).

$$P(t)=[S]_o(1-\exp(-k*t^{(1-h)})) \quad (34)$$

where P(t) is the product concentration at time t, $[S]_o$ is the initial substrate concentration, k is the reaction rate constant, and t is time.

It was thus concluded that the intact Cel7A acts in a 2-D surface phenomenon, where diffusion time would be expected to increase with increasing substrate concentration. Similarly, the action of the Cel7A core (catalytic domain) was stated to be a 3-D phenomenon since the diffusion time decreases with increasing substrate concentration.

Contrary to the classical enzyme reaction scheme, the product formation step can also be diffusion-controlled since cellobiohydrolases have to process along the cellulose chain while cleaving β-1,4-glycosidic bonds. This was incorporated in the study by Xu and Ding (2007) who derived the following equation:

$$\frac{k_2[E]t^{1-f}}{1-f} = [P] - K_m \ln\left(1-\frac{[P]}{[S]}\right) \quad (35)$$

where f is the fractal dimension, k_2 is the product formation rate constant, [E] is the enzyme concentration, [P] is the product concentration, [S] is the substrate concentration, and K_m is the Michaelis constant. The spectral dimension d_s of a bimolecular reaction is defined by $d_s = 2(1-f)$ (Kopelman, 1988). Values of f were found to be 0.44 (d_s=1.12) and 0.22 (d_s=1.56) for *T. reesei* Cel7A and *H. insolens* Cel7A respectively, implying a higher processive action for the *T. reesei* Cel7A. The effect of overcrowding of the enzymes (referred to as 'jamming') was also studied by the use of the following equation:

$$\left(1-\frac{[E]}{j[S]}\right)\frac{k_2[E]t^{1-f}}{1-f}=[P]-K_m\ln\left(1-\frac{[P]}{[S]}\right) \quad (36)$$

where j is the jamming parameter. The jamming parameter was found to be around 0.0004.

The above-mentioned two works are only semi-quantitative. They have, however, helped in understanding the role of fractal kinetics in enzymatic cellulose hydrolysis.

There is no conclusive evidence on whether enzyme diffusion on the cellulose surface is rate-limiting for the cellulose hydrolysis process or not. By measuring the diffusion rates of *Cellulomonas fimi* cellulases on *Valonia ventricosa* microcrystalline cellulose, Jervis et al. (1997) concluded that the surface diffusion of enzymes was unlikely to be rate-limiting. According to the diffusion rates measured, each cellulase traverses several hundred lattice sites in a minute. These were compared with the hydrolysis rates of *Cellulomonas fimi* endoglucanase (CenA) on bacterial microcrystalline cellulose (BMCC) – 0.23 mol glucose/enzyme/min (Meinke et al., 1993), which are lower than the diffusion rates. However, as the authors have stated, the importance of the diffusion step also depends on how the hydrolysable sites on the substrate are distributed. The substrate used in this work was highly crystalline; for other cellulosic substrates such as Avicel or Solka Floc, and those consisting of lignin and hemicellulose, it is possible that substrate heterogeneity and partial crystallinity result in rate-limiting diffusion rates. Since jamming occurs when there is overcrowding of cellulases on the substrate surface, it would be valuable to observe how the hydrolysis rates vary as the amount of adsorbed cellulase increases. Igarashi et al. (2006) measured the hydrolysis rates and specific activity of Cel7A from *Trichoderma viride* as its surface density was increased on cellulose samples from *Cladophora* and *Halocynthia*. The hydrolysis rates went through a maximum, whereas the specific activity declined continuously; this was attributed to overcrowding of enzymes on the substrate surface.

2.4. Modeling synergism of cellulase components

A mixture of cellulase components, cellobiohydrolases and endoglucanases, has higher activity than the individual components alone (Beldman et al., 1988; Fujii and Shimizu, 1986; Gusakov et al., 2007; Henrissat et al., 1985; Kleman-Leyer et al., 1996; Nidetzky et al., 1994b; Schell et al., 1999; Wood and McCrae, 1978; Woodward et al., 1988a; Woodward et al., 1988b). Modeling synergistic kinetics of the cellulases requires separate mathematical expressions for the individual components and the inclusion of cellulose chain ends as a variable in the model. The earliest of such models was proposed by Suga et al. (1975) for exo and endo-enzyme depolymerization of polysaccharides based on the Michaelis-Menten scheme. This model was extended by Okazaki and Moo-Young (1978) to include product inhibition and β-glucosidase activity. Based on these theoretical studies, DeanIII and Rollings (1992) developed a model that was inconsistent with experimental data at longer times. The following data were analyzed: conversion, polydispersity of polysaccharides, synergism, weight-averaged and number-averaged molecular weights of polysaccharides. Substrate and product inhibition, and enzyme deactivation were stated to be possible causes for the lesser predictive capability of the model at longer times. It is also possible that the model class by itself is not correct, therefore, as the authors themselves state, the above mentioned additional kinetic factors need to be incorporated in the models to ascertain the validity/invalidty.

Using substrate concentration as the only substrate variable, Fujii et al. (1981) developed a model where the endo and exo activities were represented by Michaelis-Menten expressions. The model was evaluated for carboxymethyl cellulose and hydroxylethyl cellulose (Fujii and Shimizu, 1986). Another Michaelis-Menten based model for synergism was proposed by (Nidetzky et al., 1994b) where an additional term for synergism was added to the equation:

$$v(E_1,E_2)=v(E_1)+v(E_2)+v_{syn.}(E_1,E_2) \quad (37)$$

where $v(E_1,E_2)$ is the hydrolysis rate in the presence of two enzymes E_1 and E_2, $v(E_1)$ and $v(E_2)$ are the individual hydrolysis rates, and $v_{syn.}(E_1,E_2)$ is the synergistic hydrolysis rate. However, these models based on the Michaelis-Menten scheme have limitations, as discussed in the section 2.2.2 'Michaelis-Menten based models'.

Converse and Optekar (1993) took into account enzyme adsorption, degree of polymerization, and accessibility of the substrate to model cellulose hydrolysis by cellobiohydrolase and endoglucanase. The model matched the experimental data well till a conversion level of approximately 40% (data from Woodward et al. (1988b)). The adsorption and DP variations were not, however, validated by experiments. The degree of synergism, which was shown to go through a maximum as the cellulase concentration increased, has been explained by the 'substrate inhibition' phenomenon (Väljamäe et al., 2001). At low surface coverage of the substrate (a condition achieved at high substrate concentration relative to enzyme), synergism is low as cellobiohydrolases do not benefit from the new chain ends created by endoglucanases. Substrate inhibition was also observed by Liaw and Penner (1990), and Huang and Penner (1991), but no implications of synergism were discussed. At high surface coverage (low substrate/high enzyme concentrations) competition among enzyme species for adsorption results in a decrease in synergism. Fenske et al. (1999) used Monte Carlo simulations for an enzyme that featured both endo and exo activity. Hydrolysis rates were shown to be lower at low surface coverage of the substrate due to the partial endo activity of the enzyme and went through a maximum as the substrate concentration increased. This phenomenon was termed 'auto-synergism'.

A deeper understanding of enzyme synergism is needed to optimize the mixtures of endoglucanases and cellobiohydrolases. Since the adsorbed amount of cellulases is susceptible to change along conversion, it is crucial to study these variations and their

implications on synergism. Experimental data that corroborate model predictions on variations in DP and chain size distributions are required to get accurate parameter values associated with these substrate properties. So far no work has successfully achieved such a validation. DeanIII and Rollings (1992) attempted to validate their model for non-cellulosic substrates (dextran-polysaccharide with α-1,6-glycosidic linkages) but were unable to match the experimental data at longer residence times. As the reaction proceeded, a change in the type of pattern in the size distribution was observed (Kleman-Leyer et al., 1994; Kleman-Leyer et al., 1996; Mansfield and Meder, 2003; Pala et al., 2007; Rammos et al., 1993). This shows that the susceptibility of a substrate to enzymatic attack can vary with chain size. The complexity associated with the accessibility of the available chain ends on the heterogeneous substrate is clearly a key issue that needs to be addressed before depolymerization models become informative.

2.5. Models of pure cellulosic substrates and lignocellulosic substrates

Lignin reduces the accessibility of cellulose to cellulases and also adsorbs cellulases, resulting in lower hydrolysis rates (Mansfield et al., 1999). The effect of lignin content is also evident from numerous empirical models (Table 1). Since the presence of lignin can significantly affect the hydrolysis rates, models developed for pure cellulosic substrates cannot be extended to substrates having high lignin content. For example, in the presence of lignin, a two-phase model might be applicable, whereas for pure cellulosic substrate it is not apparent. Adsorption of cellulase and β-glucosidase onto lignin has been incorporated into a few models with rate equations (Shao et al., 2009a) (see equations 12 and 13) and Langmuir isotherms (Ljunggren, 2005; Pettersson et al., 2002; Philippidis et al., 1993; Philippidis et al., 1992; Zheng et al., 2009). It was shown by Zheng et al. (2009) that their model did not match the experimental data if the negative role of lignin was ignored. Shin et al. (2006) accounted for the presence of non-cellulosic materials in steam-exploded wood by including an inhibition parameter. It has been shown that

cellulases having similar activity on pure cellulosic substrates can have different affinities for lignin (Berlin et al., 2005). Synergism results for pure cellulosic substrates might also be different for more realistic substrates since the affinity of various cellulases for non-cellulosic parts can vary. Changes in crystallinity can also be affected by lignin (Zhang and Lynd, 2004), and hence the observation of crystallinity variations along conversion must be interpreted carefully. The extent to which crystallinity limits the enzymatic conversion of biomass into sugars can depend on the lignin level and vice-versa (Zhu et al., 2008). Since lignin is not degraded by cellulases, it can act as a barrier resulting in stoppage of the enzymes on the substrate. In terms of fractal kinetics, lignin and hemicellulose act as obstacles and hence increase the fractal nature of the reaction system.

Deeper understanding of the role of lignin in enzymatic digestion of lignocellulose and its interaction with enzymes is needed not just for improving pretreatment technologies but also for engineering enzymes that have lesser affinity for lignin (Berlin et al., 2005). This is possible through quantification and modeling of lignin contribution in various steps of the hydrolysis process.

2.6 Conclusions

Cellulase hydrolysis of cellulose is a reaction in heterogeneous medium. Classical homogenous enzyme catalysis is modeled by Michaelis-Menten kinetics and heterogeneous catalysis on a catalyst support, by Langmuir-Hinshelwood kinetics. Cellulase kinetics on insoluble lignocellulosic substrates is a combination of the above two kinds of reactions and also involves other factors (product inhibition, enzyme deactivation, substrate crystallinity, substrate accessibility changes, substrate reactivity changes, fractal nature of the reaction, changes in enzyme synergism, lignin inhibition), which result in retarding the rates at higher degrees of conversion. While the models in literature have not pinpointed the exact mechanism of enzymatic action on lignocellulosic

materials, they have helped in understanding the various factors that are at play. Additional insight will be made possible by models consisting of the major substrate and enzyme properties (substrate – concentration, DP, accessible fraction, size-distribution of chains, crystallinity; enzyme – individual component concentration, synergistic/competitive factors, and adsorbed concentration of individual components). However, due to the increase in the number of parameters, overparameterization resulting in unidentifiable parameters can be an issue (Sin et al., 2009). With the improvements in measurement techniques likes fluorescent detection of enzyme generated reducing ends (Kurasin and Valjamae, 2011), cellulose crystallinity determination methods (Bansal et al., 2010; Barnette et al., 2011; Park et al., 2009), and observation of cellulase movement on cellulose surface (Igarashi et al., 2009), this problem can be overcome. It is clear from the research reviewed in this article that adsorption, substrate reactivity, and accessibility can change along conversion. Therefore, their dynamic nature must be taken into consideration when building models. The range of conversion for checking the predictive ability of a model is also important, since major slowdowns are observed at high conversions. Only one-third of the models reported have been validated with data beyond 70% conversion (Table 1).

Improvements in enzyme catalysis have mainly been guided by the engineering of the active site or amino acid residues identified as playing an important role. In the case of cellulases and their kinetics on insoluble lignocellulosic substrates, rate limitations cannot be explained solely by active-site considerations, mostly because of the heterogeneity of the substrate. Information regarding the catalytic domain, the binding domain, and the linker region (the three domains of a cellulase) through advances in structural biology will certainly contribute to a more complete understanding of the operation of cellulases at the molecular level. Additionally, to significantly improve the enzymatic process, contributions of the various substrate characteristics need to be quantified to specifically target the enzyme and substrate features that need improvement.

CHAPTER 3

ELUCIDATION OF CELLULOSE ACCESSIBILITY, HYDROLYSABILITY AND REACTIVITY AS MAJOR LIMITATIONS IN THE ENZYMATIC HYDROLYSIS OF CELLULOSE

(Experimental work associated with this chapter was carried out by Prabuddha Bansal and Bryan Vowell)

3.1. Introduction

Many hypotheses for the rapid decline in the rates of cellulose biohydrolysis have been proposed ((Bansal et al., 2009), Chapter 2) – enzyme inactivation (Caminal et al., 1985; Converse et al., 1988), changes in substrate accessibility and reactivity (Hong et al., 2007; Kumar and Wyman, 2009), increase in cellulose crystallinity (Chen et al., 2007), depletion of cellulose chain ends for cellobiohydrolases (Hong et al., 2007), decrease in enzyme synergism, surface obstacles (Jalak and Valjamae, 2010; Kurasin and Valjamae, 2011) and fractal nature of the substrate (Väljamäe et al., 2003; Xu and Ding, 2007). However, determination of the key factors and their quantification has remained experimentally challenging. Mechanistic understanding of the rate-limiting causes, and changes in cellulose-cellulase interactions with reaction time is further confounded by conflicting reports on evolution of parameters such as crystallinity, reactivity, and accessibility during the course of the reaction (Bansal et al., 2009; Kumar and Wyman, 2009; Lynd et al., 2002).

In this thesis, results from both computational and experimental studies are used to sift through diverse hypotheses on rate limitations, and identify as well as quantify the major ones. Cellobiose product inhibition, which can be alleviated using an excess of β-

glucosidase (Bommarius et al., 2008), was not considered in this thesis. Using model-guided experiments, changes in accessibility (substrate available for cellulase adsorption), reactivity (hydrolytic activity per amount of actively adsorbed cellulase), and hydrolysability (reactive/hydrolysable fraction of accessible cellulose resulting in productive adsorption) were quantified on a pure cellulosic substrate (Avicel PH-101). Unproductive adsorption, which can be due to various reasons (lack of reactive sites, improper orientation of the cellulose chain with respect to the catalytic domain, inaccessibility of chain ends (Kongruang et al., 2004), substrate competition between the adsorbed enzymes, jamming (Bommarius et al., 2008), obstacles (Jalak and Valjamae, 2010; Kurasin and Valjamae, 2011)), was hypothesized to be giving rise to the non-hydrolysable fraction of the cellulose.

Reactivity is a term that has been used broadly in two different contexts – i) in kinetic models to explain the reduced digestibility of hydrolyzed cellulose (Bansal et al., 2009), and ii) in restart experiments (where enzymes are washed off the surface of unreacted cellulose and the partially hydrolyzed substrate is subjected to cellulase hydrolysis under initial conditions) to study the rate of hydrolysis of partially converted cellulose at a chosen substrate and enzyme concentration (Desai and Converse, 1997; Drissen et al., 2007; Gusakov et al., 1985; Hong et al., 2007; Kumar and Wyman, 2009; Lee and Fan, 1983; Ooshima et al., 1991; Väljamäe et al., 1998; Yang et al., 2006; Zhang et al., 1999). Since the enzymatic hydrolysis of cellulose is a reaction involving cellulose as well as the enzyme, in this thesis the term ‘reactivity’ has been used to quantify the rate of hydrolysis for a productively bound cellulase on cellulose.

Overparameterization of models, which happens when only time conversion data is used for estimating parameters (Sin et al., 2009), was avoided by determining the respective parameters through independent restart experiments at various conversion levels.

3.2. Materials and methods

Materials

All chemicals and reagents were purchased from Sigma (St. Louis, MO, USA) unless otherwise stated. Avicel PH-101, cellulase cocktail from *T. reesei* (159 FPU/mL), and β-glucosidase (from almonds, 5.2 U/mg) were obtained from Sigma and phosphoric acid (85%) was obtained from EMD (Gibbstown, NJ, USA). *Trichoderma reesei* QM9414 strain was obtained from ATCC (#26921; American Type Culture Collection, Manassas, VA, USA). The BCA protein assay kit was obtained from Thermo Fischer Scientific (Rockford, IL, USA).

Cellulose hydrolysis

Cellulose (20 mg/mL) (Avicel or partially converted Avicel) was added to sodium acetate buffer (1 mL, 50 mM, pH 5) and allowed to hydrate for 1 hour at 50 °C and 900 rpm. The hydrolysis was initiated by the addition of β-glucosidase and cellulases and stopped by centrifugation at 4 °C, 14000 rpm. The cellulase concentration was varied from 13.8 μg/mg cellulose to 640 μg/mg cellulose. To prevent cellobiose product inhibition, cellulase/β-glucosidase activity ratio was kept at 1:20 (Bommarius et al., 2008).

Determination of glucose content

Glucose concentration was determined by way of the DNS (dinitrosalicylic acid) assay, as described previously (Bommarius et al., 2008). The DNS assay was compared with HPLC analysis and found to yield identical results.

Enzyme adsorption study

Cellulose (20 mg/mL) was added to sodium acetate buffer (1 mL, 50 mM, pH 5) and allowed to hydrate for 1 hour at 50 °C and 900 rpm and the mixture was then cooled down to 4°C. Cellulases were added in various amounts and the mixture was further

agitated for 30 min at 4 °C. After centrifugation, the supernatant was collected and protein content analysis was performed using the BCA protein assay. In order to eliminate interference from agents other than proteins (e.g. glucose and salts, etc.), proteins were selectively precipitated using deoxycholate and trichloroacetic acid (Brown et al., 1989), collected and re-suspended prior to using the BCA protein assay kit.

Partially converted Avicel

Samples with an initial cellulose concentration of 20 mg/mL were subjected to the hydrolysis procedure described above for a period of time required to reach the desired level of conversion and partially converted cellulose was collected by centrifugation. An enzyme desorption procedure was then used to purify the cellulose of bound proteins and the cellulose samples were subsequently freeze-dried.

Enzyme desorption

The enzyme desorption procedure was adapted from a previously published method (Hong et al., 2007). Cellulose samples obtained after partial conversion were suspended in a solution of 1.1% SDS in water and incubated at 80 °C (water bath) for 15 minutes. The samples were then washed 3 times with 75% ethanol and four times with water. Control samples were prepared which consisted of buffer-suspended cellulose without the addition of protein. These samples were subjected to the enzyme desorption procedure as well as lyophilization. These steps were found to have little to no effect on adsorption and hydrolysis behavior (Appendix A), and crystallinity.

Instantaneous rates

For calculating the instantaneous rates used in section 3.3.1, an empirical curve (equation 1) (Väljamäe et al., 2003) was fitted to the conversion curves of Avicel and phosphoric

acid swollen cellulose (PASC, generated by Dr. Mélanie Hall as previously published (Bansal et al., 2010; Hall et al., 2010a)) hydrolysis for 160 μg cellulase/ mg cellulose.

$$X = 1 - \exp(-kt^{(1-h)}) \quad (1)$$

where X is the conversion level, and k and h are fitted parameters to the curve.

For Cel7A hydrolysis, protein concentration was 40 μg of purified enzyme per mg of Avicel.

Cel7A purification, determination of chain ends per amount of substrate, and hydrolysis with pure Cel7A was carried out by Dr. Mélanie Hall. Description of Cel7A purification, and chain end determination can be found in Appendix B.

3.3. Results and discussion

3.3.1 Rate order and cellulose crystallinity

Decreasing rates at high conversion, though suggested to be caused by multiple factors (see Introduction), may potentially simply be also the result of substrate depletion. The rate of cellulose hydrolysis by enzymes can be expressed by equation (2).

$$\text{Rate} = dX/dt = k*S*[E]_{ads}*f \quad (2)$$

where X is the conversion level, S is the substrate concentration, $[E]_{ads}$ is the concentration of cellulases adsorbed per amount of substrate, f is the fraction of productively adsorbed cellulases, and k is the reaction rate constant (reactivity) reflective of the speed of a processive cellobiohydrolase or the rate of hydrolysis by an endoglucanase.

Equation (2) can also be written as equation (3), where instantaneous substrate concentration is expressed in terms of the initial substrate concentration.

$$\text{Rate} = dX/dt = k*S_0(1-X)*[E]_{ads}*f \quad (3)$$

where S_0 is the initial substrate concentration.

In the absence of any rate hindrances, k will be a constant, f will equal unity, and $[E]_{ads}$ will be at the maximum adsorbable capacity. Equation 3 then simply decomposes into the following form:

$$\text{Rate} = dX/dt = C*(1-X) \quad (4)$$

which is a simple first-order rate expression. For any n^{th} order rate expression $\text{Rate} = C*(1-X)^n$, a plot of ln(rate) vs. ln(1-X) will give a slope of n.

These plots for amorphous (PASC) and crystalline cellulose (Avicel – 60% crystalline (Hall et al., 2010a)) are shown in Figure 4. The apparent rate order for PASC was found to be close to 1, showing an absence of any significant rate hindrance. For Avicel, however, the rate order changed over time. Two distinct phases could be identified with rate orders of close to 8 in the initial stages (up to a conversion level of about 23%), and almost 2 in the latter stages up to a conversion of 92%. These observations clearly point to involvement of rate-limiting factors and deviations from the ideal situation of a first-order reaction rate. Moreover, changing rate orders are characteristic of fractal kinetics (Anacker and Kopelman, 1987; Kopelman, 1988; Kopelman, 1986) which occurs in spatially constrained media, and/or in the presence of obstacles mixed non-homogeneously in the reaction system. Since crystalline cellulose is insoluble in solution, the enzymes have to adsorb on to the substrate and hydrolyze cellulose in a one-

dimensional system along the cellulose fiber (for cellobiohydrolases, this action is processive in nature along a cellulose strand (Divne et al., 1994; Divne et al., 1998)).

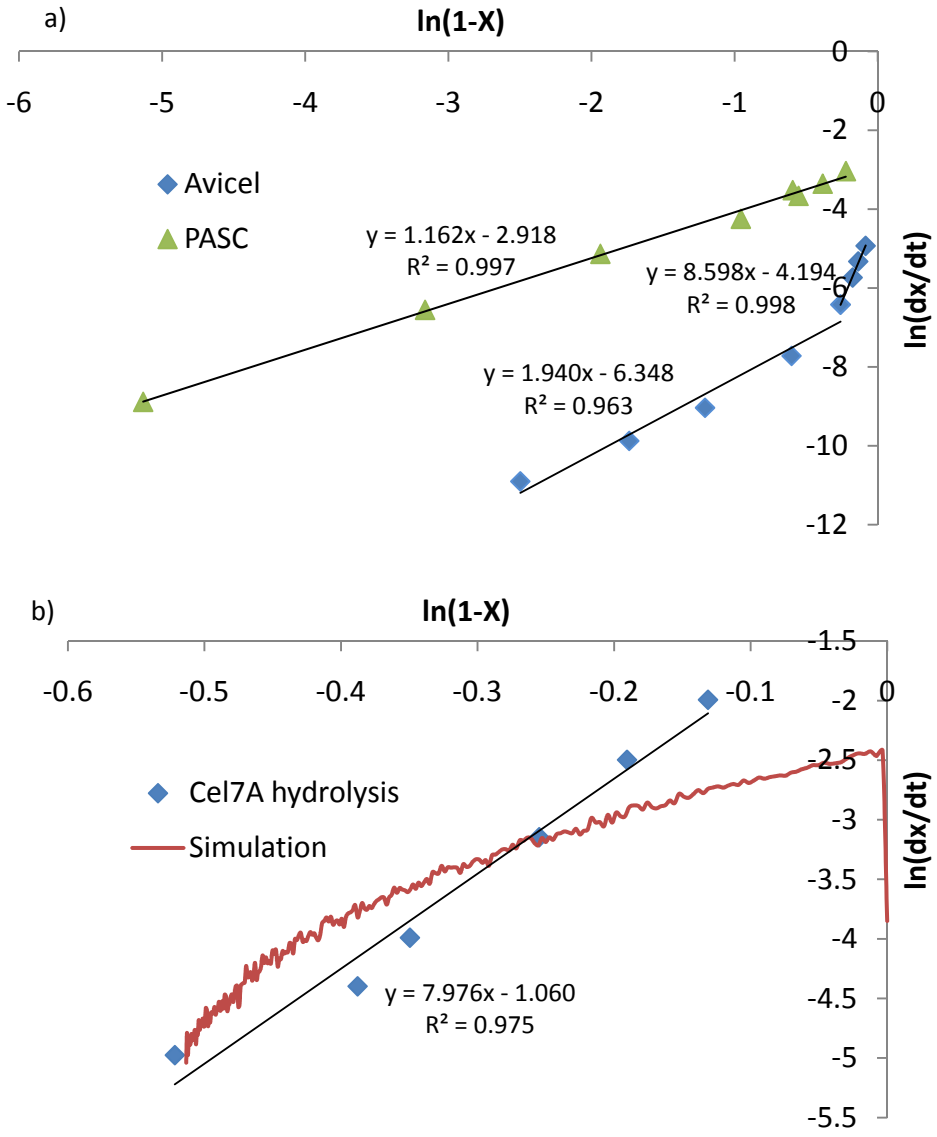


Figure 4. Levenspiel plot (Levenspiel, 1999) (ln(dX/dt) vs. ln(1-X)) for a) Avicel and PASC with cellulase mixture, and b) Avicel hydrolysis with pure Cel7A and simulations with cellobiohydrolase. X – conversion, t – time.

3.3.2 Micro-kinetic simulation to evaluate enzyme clogging as a first-order phenomenon

Recent works have shown that cellobiohydrolases possibly get stuck at obstacles after the first few hydrolytic cycles (Jalak and Valjamae, 2010; Kurasin and Valjamae, 2011). Here, a stochastic model was developed to determine whether this proposed hypothesis was a first-order based phenomenon. The simulation was limited only to cellobiohydrolases by constraining the cellulase action from one chain end only (in the simulations, only one end of the chains was reactive, analogous to reducing ends for cellobiohydrolases I). A simplified schematic of the model is shown in Figure 5. The model parameters are t_f – time to find chain end, P_c – probability of clogging, t_p – time for hydrolysis (cleavage of a β-glycosidic bond), and P_p – probability of processing along a cellulose strand.

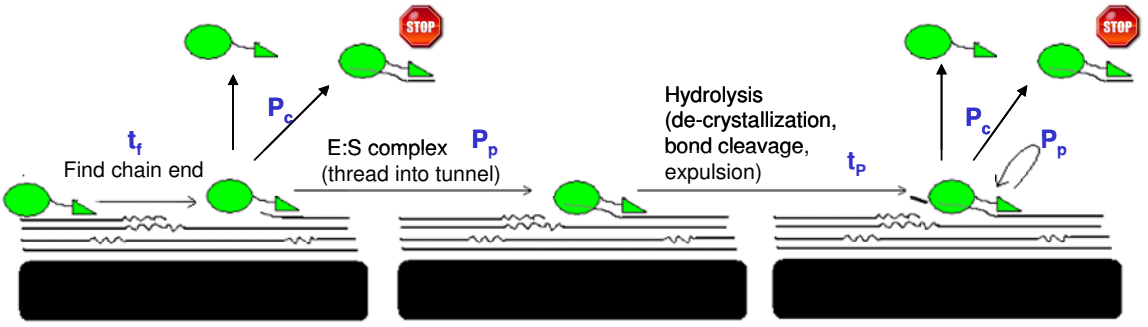


Figure 5. The stochastic model. Model parameters were set to: $t_f = 5$, $t_p = 1$, $P_c = 0.001$, $P_p = 0.8$.

The apparent rate-order obtained from the simulations in MATLAB ®(The Mathworks Inc. R2008b) was found to be different from that observed experimentally (Figure 4 b), the convex nature of the curve indicating a faster slow-down. As mentioned earlier,

changing rate orders are characteristic of fractal kinetics. Therefore, difference in trends of rate orders implies different fractal kinetics at a microscopic level. Modeling of enzyme clogging as a first-order event is evidently not an ideal explanation of the rate hindrance. Evolution of substrate morphology along conversion, possibly akin to a substrate erosion phenomenon (Väljamäe et al., 1998), could be governing the rate at which enzymes become inactive due to clogging. Recently, Praestgaard et al. (2011) studied the initial 'burst phase' analytically, and found that enzymes getting stuck at 'check blocks' could not account for the decreasing rates except at initial stages of the hydrolysis.

When the model parameters were fit to time-conversion data, a degenerate set of parameters was found to fit the data (more than one set of t_f, t_p, P_c, and P_p fit the data). Such overparameterization of the model can be understood from the fact that the hydrolysis rates depend only on the enzymes in the active state, and this can be controlled by varying the enzymes either in the clogged or free state, thereby giving rise to two degrees of freedom. This is a short-coming of many models in the literature (Sin et al., 2009). Nevertheless, qualitative comparisons can still be made, as has been done with the apparent rate orders.

3.3.3 Change in cellulose crystallinity and degree of polymerization along conversion

Cellulose crystallinity has been investigated in many works, and is known to be a major rate-governing property (Bansal et al., 2010; Hall et al., 2010a; Lynd et al., 2002; Mansfield et al., 1999; Zhang and Lynd, 2004). However, there are conflicting reports regarding its increase or decrease along conversion (Bansal et al., 2009), and therefore, the role it plays in controlling the rates. Partial reason for the lack of consensus lies in the differences in the measurement techniques used to calculate cellulose crystallinity, which themselves impart a lot of error in the crystallinity index numbers (Park et al., 2010).

When partially converted cellulose samples were recovered from the reaction mixture, the crystallinity index determined by solid state ^{13}C-NMR (Hall et al., 2010a) as well as X-ray diffraction (with crystallinity calculation method (Bansal et al., 2010)) (Figure 6) was found to remain constant. Therefore, though crucial as rate-determining factor *before* the reaction, cellulose crystallinity is not a parameter that needs to be taken into account during the hydrolysis reaction itself.

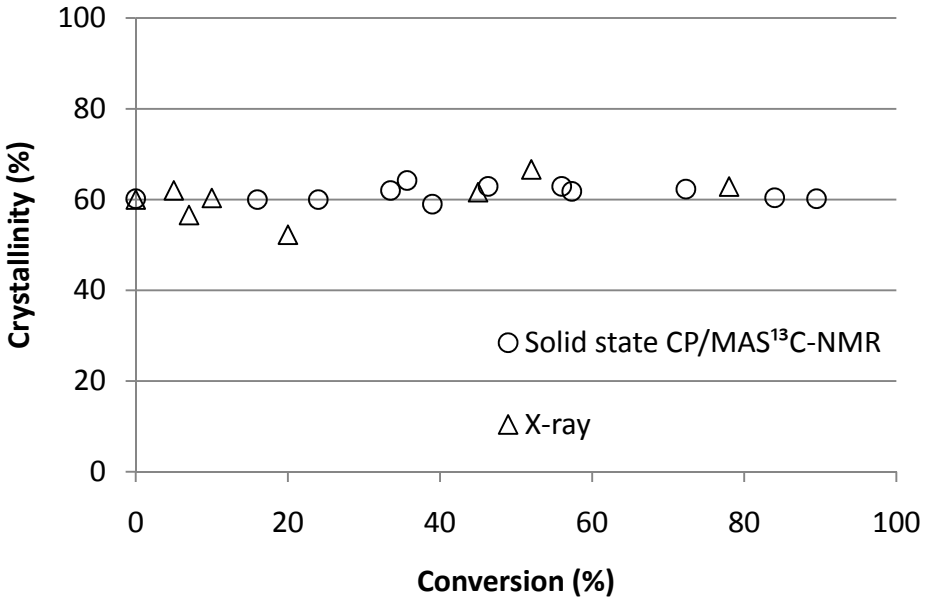


Figure 6. Cellulose crystallinity along conversion. Avicel® hydrolysis conditions: Cellulase/β-glucosidase 1:20 activity ratio, 20g/L cellulose, 50mM NaOAc buffer pH 5.0, 50 °C.

The number of chain ends per amount of substrate is simply the inverse of the number-average degree of polymerization. This can be a critical factor determining the rate as cellobiohydrolases act specifically from chain ends. The number of chain ends per amount of cellulose was not found to vary significantly with conversion (17 – 18

nmol/mg cellulose; experiments by Dr. Mélanie Hall). However, it is currently extremely challenging to experimentally determine what fraction of the chains ends is accessible to cellobiohydrolases, and whether this fraction changes along conversion. The fraction of chain ends accessible to solvent has been shown to be 60% for Avicel (Kongruang et al., 2004), but this fraction for cellobiohydrolases has not been quantified yet.

3.3.4 Macro-kinetic studies to identify rate limitations

The factors responsible for rate retardation are identifiable from the rate expression in equation 3.

$$\text{Rate} = dX/dt = k*S_0(1-X)*[E]_{ads}*f \quad (3)$$

Other than increase in the conversion X, the factors k, $[E]_{ads}$, and f can also change during the reaction causing a decrease in the rate. The rate constant k for the β-glycosidic bond cleavage reflects the intrinsic reactivity of the substrate. $[E]_{ads}$ is the total amount of enzyme adsorbed per amount of substrate, and is related to the accessibility of the substrate (accessible fraction – substrate available for cellulase adsorption). The fraction of adsorbed enzymes in an active state, f, will be determined by the hydrolysable fraction of the accessible substrate.

Concept of the substrate is shown in Figure 7.

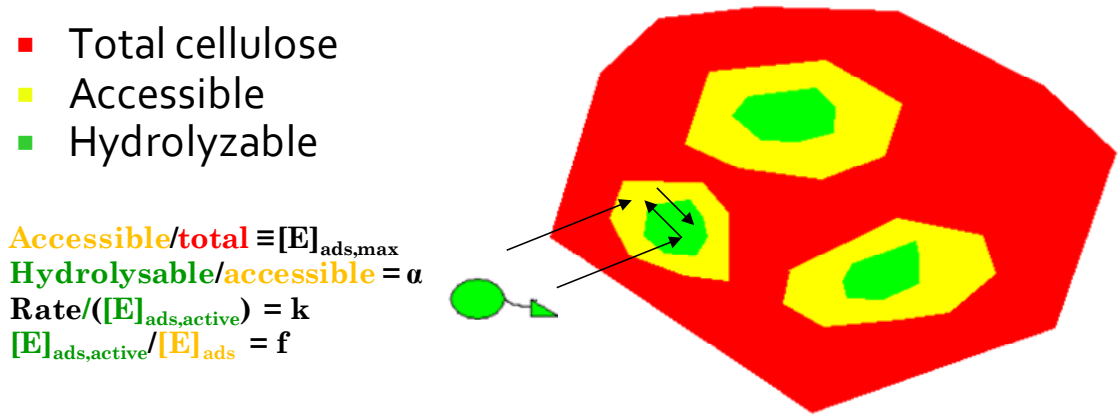


Figure 7. Concept of cellulose substrate with total cellulose (red), accessible (yellow), and hydrolysable portions (green). Arrows indicate parts of the substrate, onto which the cellulases can adsorb, and between which they can change states.

During hydrolysis, the enzymes adsorb on to either the hydrolysable or non-hydrolysable (but accessible) part of the substrate, and change states between the two *via* diffusion (Jervis et al., 1997) on the substrate (with no hydrolysis). As cellulases hydrolyze cellulose, the ratio of accessible to total (yellow plus green to red compared to red, Figure 7), and hydrolysable to accessible (green compared to green plus yellow) can decrease. The accessible fraction of cellulose at various conversion levels can be determined experimentally by measuring its maximum adsorption capacity. At various adsorbed quantities, the initial hydrolysis rates (defined here as amount of glucose produced in 10 minutes) are expected to increase until full coverage of all productive adsorption sites (the hydrolysable portions), followed by a saturation phase resulting from non-productive adsorption (no further increase in reaction rate).

Adsorption can be unproductive due to various reasons – lack of reactive sites, improper orientation of the cellulose chain with respect to the catalytic domain, or substrate competition between the adsorbed enzymes. Even though we cannot explicitly

measure or determine the cause, we can relate productive adsorption to hydrolysability. The equilibrium reaction in equation 5 eventually yields a correlation between hydrolysability and productive adsorption (equation 8). The stochastic simulation studies mentioned earlier showed that even when the time to find a chain end for a CBH was one to two orders of magnitude higher than the hydrolysis time, the concentrations of active and inactively adsorbed enzymes quickly reached a steady equilibrium (data not shown).

$$\underset{\substack{\text{Adsorbed, chain}\\ \text{not in catalytic domain}}}{E_a/mS} + \underset{\substack{\text{Hydrolyzable substrate,}\\ \text{Unoccupied}}}{mS_{hyd,free}} \rightleftharpoons \underset{\substack{\text{Productively bound}\\ \text{enzyme}}}{E_a{:}mS} + \underset{\substack{\text{Accessible substrate,}\\ \text{unoccupied}}}{mS_{acc,\ free}} \tag{5}$$

$$\frac{[E_a{:}mS]}{[E_a/mS]} = K^* \frac{[S_{hyd,free}]}{[S_{acc,\ free}]} = K^* \frac{(\alpha[S]_{acc,\ total} - m[E_a{:}mS][S]_{total})}{([S]_{acc,\ total} - m[E]_{ads}[S]_{total})} \tag{6}$$

(Enzyme species' concentrations in equation 6 are per amount of substrate, S)

where K is the equilibrium constant for the reaction in equation 5, m is the number of glucose sites covered by an adsorbed cellulase. Dividing the numerator and denominator on the LHS by $[E]_{ads}$ (the adsorbed cellulase concentration), and the numerator and denominator on the RHS by $[S]_{acc,\ total}$ (= $m^*[E]_{ads,max}^*[S]_{total}$), we obtain -:

$$\frac{f}{1\text{-}f} = K^* \frac{(\alpha - f^*y)}{(1-y)} \tag{7}$$

where $y = [E]_{ads}/[E]_{ads,\ max}$ is the fractional adsorption, $\boldsymbol{\alpha} = [S]_{hydrolyzable}/[S]_{accessible}$ is the hydrolysability, $f = [E{:}mS]/[E]_{ads}$ is the fraction of actively adsorbed cellulases. Note that when K >> 1, the above equation is quadratic in f, and simplifies to two roots:

$f = 1$, or $f = \alpha/y$ (8)

With K>>1, any unproductively bound enzyme will find a reactive site within a short time span (f = 1), and when all the productive sites are covered (hydrolysable portions in Figure 7), subsequent adsorption will not result in any further hydrolysis (f = α/y). Although K need not be much greater than 1, as will be shown later the biphasic behavior of hydrolysis rates vs. enzyme adsorbed will validate this assumption. Equation 5 does not assume any obstacles or clogging sites on the substrate, and may hold only for the initial phase of the hydrolysis, which for the purposes of the experimental design of this thesis is applicable, because restart hydrolysis studies were used to quantify reactivity and hydrolysability.

Based on the above explanation, kinetic studies have been conducted following the design shown in Figure 8.

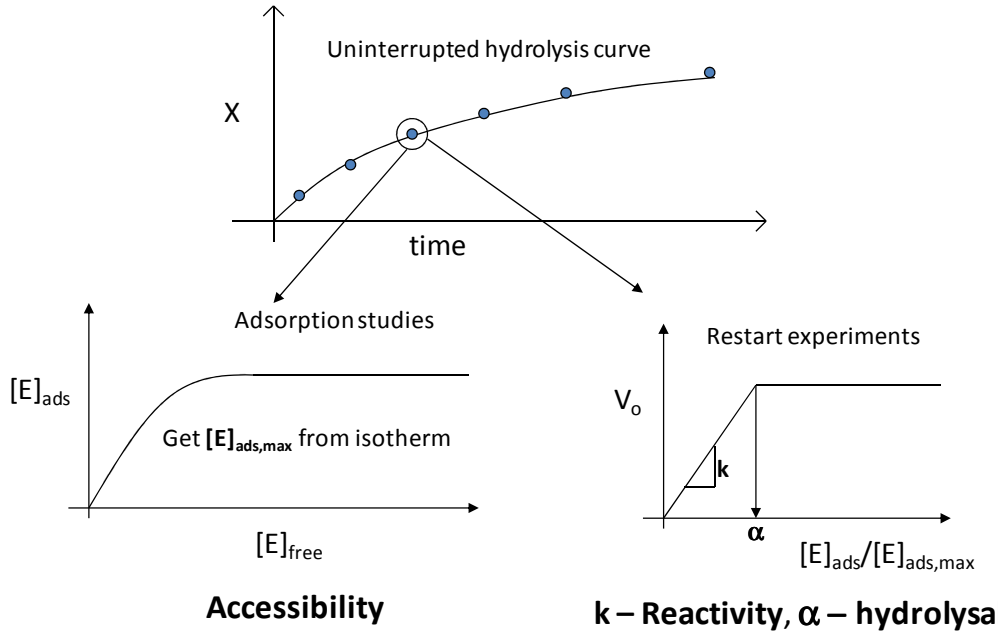


Figure 8. Experimental design: enzymes are washed off at a chosen conversion level, and adsorption studies and restart experiments are conducted to determine the changes in accessibility, reactivity, and hydrolysability. V_o is the initial rate measured in terms of glucose produced in 10 minutes.

The linear phase in the plot of V_o vs $[E]_{ads}$ also becomes apparent from equation 9, where at low enzyme concentrations and short time rates, the rate is proportional to the adsorbed enzyme concentration. The three parameters of interest are associated with accessibility ($[E]_{ads,max}$), reactivity (k), and hydrolysability (α).

$$\text{Rate} = k*S*[E]_{ads}*f = k*S_0*[E]_{ads} \qquad (\text{when } f = 1,\ S = S_0) \qquad (9)$$

3.3.5 Accessibility

Subsequent to removal of cellulases, partially converted cellulose (12%, 31%, 48%, and 66%) was subjected to adsorption studies. The adsorption data was fit to the Langmuir isotherm (Table 2, Figure 9 b). Even though the underlying assumptions of the Langmuir isotherm (reversibility, non-interacting adsorbed species, homogenous binding sites and uniform composition of adsorbed cellulase mixture) may not be valid at 4°C, it can still be used to determine the maximum adsorption capacity. Results in Figure 9 show a steady decline in the adsorption capacity with conversion.

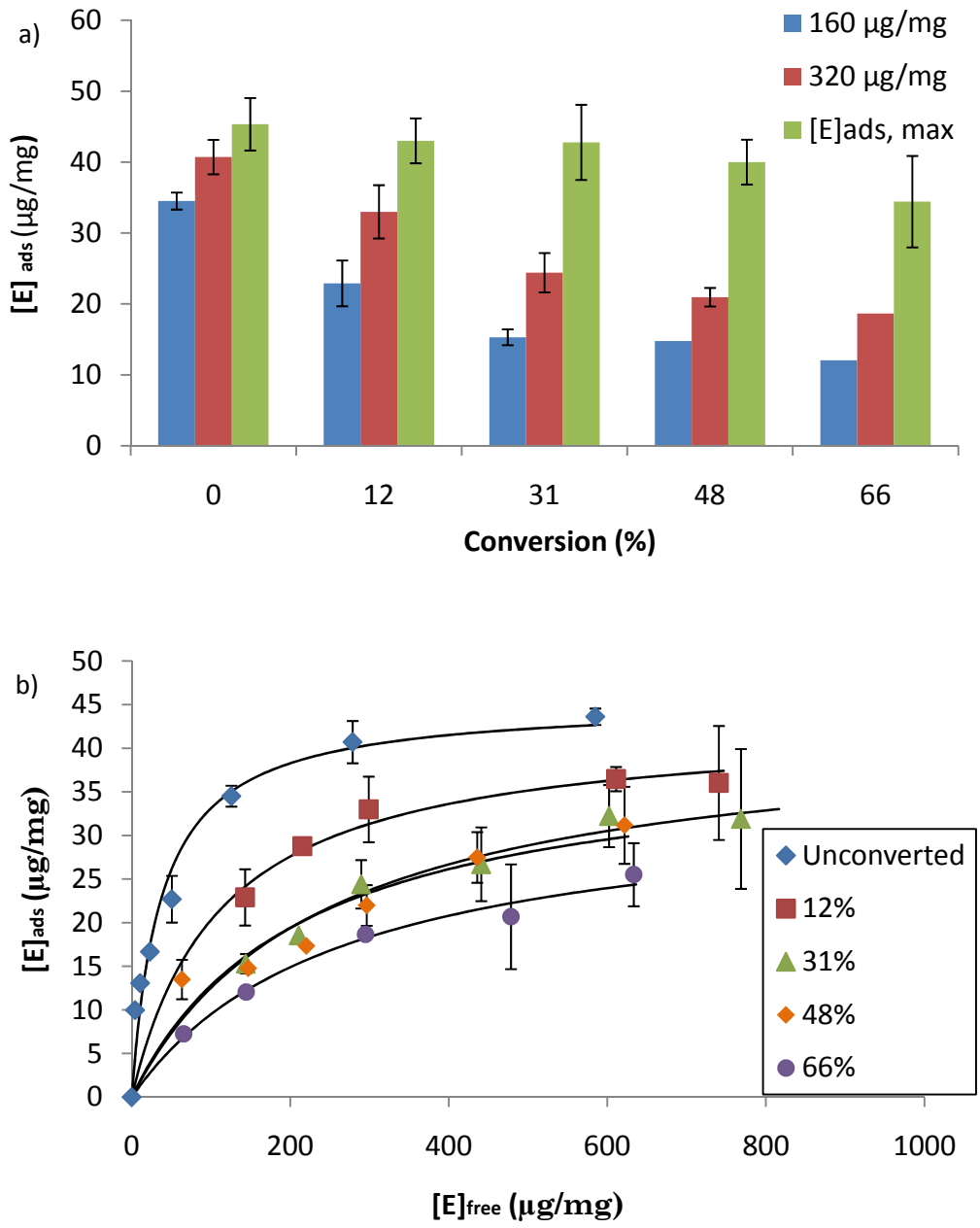


Figure 9. a) Maximum enzyme adsorption capacity ($[E]_{ads,\ max}$), and adsorbed cellulase for enzyme loadings of 160 µg/mg and 320 µg/mg, b) Adsorption data (symbols) and fitted isotherms (solid lines) for various conversion levels.

These results are in agreement with (Hong et al., 2007) who also reported a decrease in the adsorption capacity with conversion. While the adsorbed cellulase quantities decrease

steadily with conversion at various enzyme loadings (below saturation), the fitted parameter $[E]_{ads,max}$ does not show a very strong trend (Figure 9 a, Table 2). Though this could be an artifact of the statistical fitting of the Langmuir isotherm parameters as seen with the comparatively large standard deviations in $[E]_{ads,max}$, the monotonic decrease in the adsorption capacity measured with good repeatability (Figure 9 a) is clear evidence of a continuous decrease in accessibility of cellulose with conversion and change in substrate morphology. Even though $[E]_{ads,\ max}$ shows a slight downwards trend, K_{ad} (adsorption equilibrium constant) shows a marked decline with conversion (Table 2, Figure 13), and is majorly responsible for the decreasing accessibility. The relatively large standard deviations in $[E]_{ads,max}$, and K_{ad} are probably due to fitting to data at only five enzyme loadings (it is experimentally difficult to obtain reproducible protein concentration measurements at very high and very low enzyme loadings).

Yang et al. (2006) and recently Kumar and Wyman (2009) found $[E]_{ads,\ max}$ for Avicel to remain constant with conversion. The fitted $[E]_{ads,max}$ values in this thesis also do not vary strongly, so we may be driven to conclude that accessibility does not change drastically. However, as can be seen clearly for the enzyme loadings studied, there is a clear decrease in cellulase adsorption, so the results on both $[E]_{ads,max}$ and K_{ad} have to be considered to judge adsorption as a function of conversion.

Table 2. Langmuir isotherm[a] parameters and R^2 of the statistical fit.

Conversion (%)	$[E]_{ads,max}$ (µg/mg)	K_{ad} ($*10^3$) $(\mu g/mg)^{-1}$	R^2
0	45.34	27.10	0.954
12	43.00	8.94	0.952
31	42.79	4.14	0.969
48	40.00	4.72	0.921
66	34.42	3.82	0.980

[a]Langmuir isotherm: $[E]_{ads} = K_{ad}[E]_{ads,max}[E]_{free}/(1+K_{ad}[E]_{free})$. R^2 was calculated by comparing the predicted adsorbed concentrations with the mean adsorbed concentrations. Parameters estimated through non-linear parameter estimation toolbox (nlinfit) in MATLAB® (The Mathworks Inc. R2008b).

3.3.6 Reactivity

Substrate reactivity was investigated by hydrolyzing the partially converted cellulose (free of enzymes) at the starting substrate concentration (20 mg/mL cellulose) and at different enzyme loadings (Figure 10). Reactivity is the hydrolysis rate per amount of productively adsorbed cellulase. It is also an estimate of the rate constant k in equation (3). Glucose produced in 10 minutes was taken as the reference for initial rate; for unconverted Avicel, trends were very similar to 5 and 20 minute hydrolysis (data not shown).

The decrease in rate as a function of conversion observed at various enzyme loadings (Figure 10 a&b) can either be due to a reduction in the accessibility $[E]_{ads,max}$, reactivity k, or the fraction f, captured in hydrolysability α. No significant change in reactivity was observed, as shown by the overlap of the hydrolysis data in the linear regime at different enzyme loadings (Figure 10 c); the slope of the beginning linear phase is an estimate of k (equation 9). The decrease in rate observed (Figure 10 a&b) at different enzyme loadings is therefore due to a decrease in accessibility and hydrolysability as seen in the leveling off of the hydrolysis rates with increase in adsorption (see section 3.3.7).

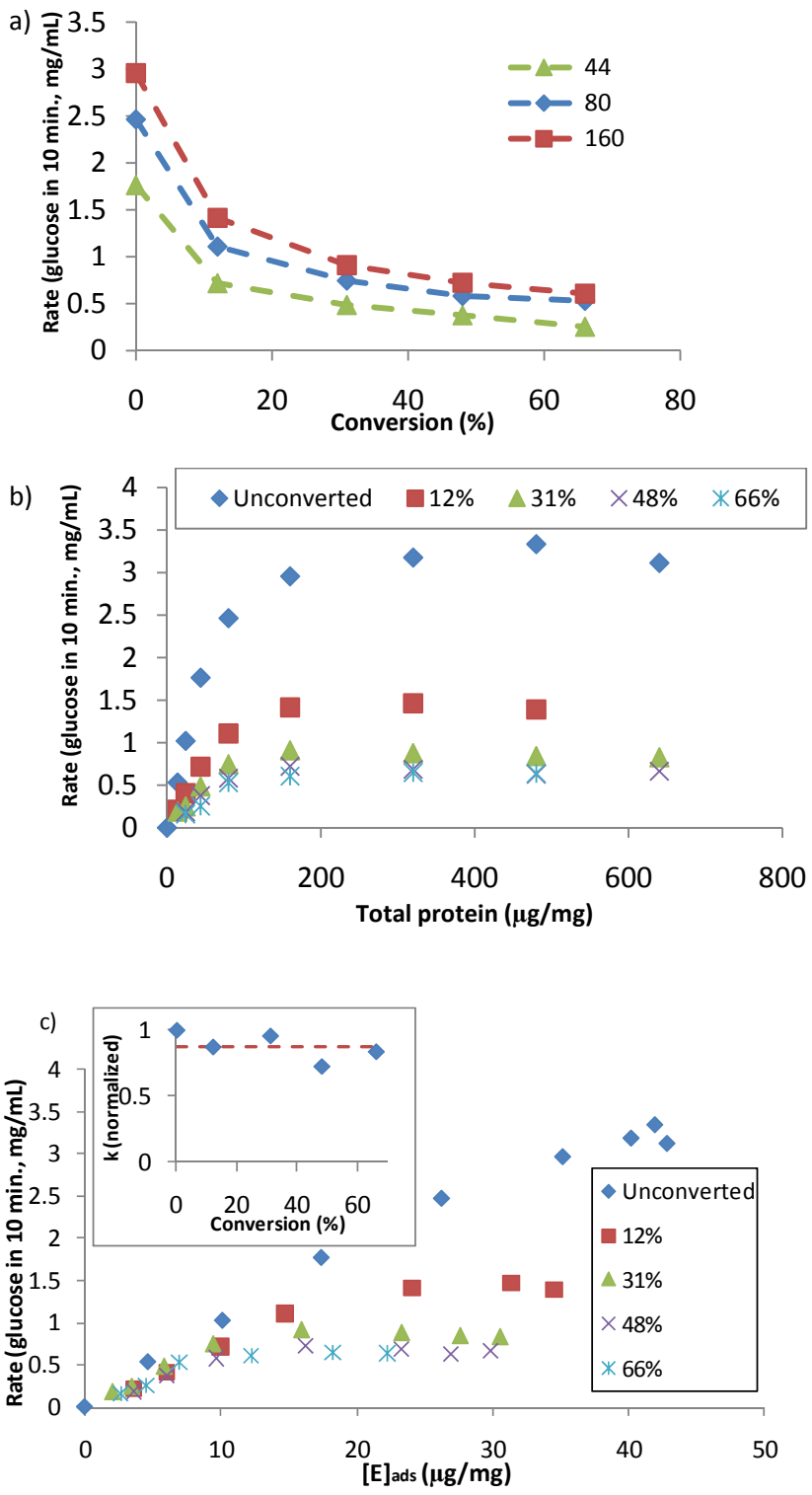


Figure 10. a) Restart rates vs. conversion levels for three enzyme loadings – 44 (▲), 80 (♦), and 160 (■) µg/mg, b) Restart rates for various enzyme loadings, c) Restart rates as a function of adsorbed enzyme concentration for various conversion levels, inset shows normalized k as a function of conversion.

Probing substrate reactivity using restart experiments has been the focus of many works, but there is no consensus regarding the decline of reactivity (Bansal et al., 2009; Kumar and Wyman, 2009; Lynd et al., 2002). Differences in the type of substrates, enzymes, and material handling methods (e.g. desorption procedures) can be one source of dissonance. The other cause of this inconsistency is the definition of reactivity itself, or rather, the experimental quantity used to estimate it. For example, Figure 10 b could point at a decrease in reactivity, but Figure 10 c shows that the amount reacted per amount of actively adsorbed cellulase is nearly the same. The fraction of productively bound enzymes can, however, decrease with conversion, giving rise to lower rates.

3.3.7 Hydrolysability

An interesting feature of the hydrolysis trend from Figure 10 c is the saturation after the linear increase. This is due to the exhaustion of the hydrolysable sites. Part of the decrease in rate observed at high enzyme loadings at a given conversion level is attributed to unproductive binding resulting from saturation of the reactive sites. The hydrolysability α, calculated as the fraction of accessible substrate that is reactive, was obtained from the restart experiments (Figure 8 & Figure 10). α was found to decrease by about 75% till a conversion level of 31%, beyond which it remains constant (Figure 11). This sharp decline in hydrolysability is consistent with the picture of a decreasing ratio of hydrolysable to accessible substrate (Figure 7). The fraction of productively adsorbed enzymes, f in equation 3 and 8, also followed a similar trend, and the decrease was sharper for higher enzyme loadings (Figure 11).

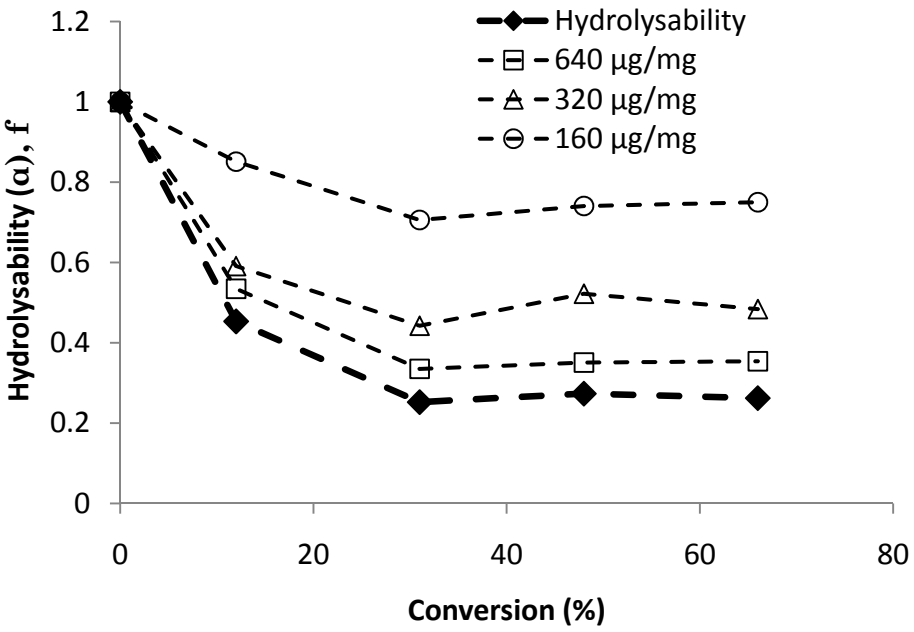


Figure 11. Hydrolysability α for various conversion levels, and f for three enzyme loadings. α was calculated as $([G]_{sat}/k)/[E]_{ads,max}$, where $[G]_{sat}$ is the saturation rate (glucose in 10 minutes, mg/mL). f was calculated as $\alpha/y = ([G]_{sat}/k)/[E]_{ads}$.

3.3.8 Accounting for rate retardation and quantification of blocked/clogged cellulases

To quantify the fraction of rate retardation accounted for by the three factors mentioned above (accessibility $[E]_{ads,max}$, reactivity k, and hydrolysability α), the predicted rates at the studied conversion levels (12%, 31%, 48%, 66%) were compared with rates from uninterrupted hydrolysis reaction. Rates were calculated by equation 3 using estimated parameter values of k, f, and the measured values of $[E]_{ads}$. Overall, the trend in restart rates seems to follow that of the uninterrupted experiment (Figure 12; the restart hydrolysis in this case was performed with a concentration equivalent to 20*(1-X) mg/mL to match the substrate concentration under actual hydrolysis conditions). However, a difference between the restart rates and the uninterrupted rates is observed (fairly constant from 10 to 66% conversion at ~ 0.3 mg/mL), that is also predicted by the

restart model for partially converted cellulose. This difference becomes quite significant at 66% conversion where the predicted hydrolysis rate is 8.5 times that of the uninterrupted one (0.34 mg/mL vs. 0.04 mg/mL).

The observation of restart rates being higher than the uninterrupted ones points to an enzyme clogging phenomenon (erosion model) (Jalak and Valjamae, 2010; Kurasin and Valjamae, 2011; Väljamäe et al., 1998). The restart kinetic study model was derived under the assumptions of no clogging. When the enzymes are allowed to re-adsorb onto partially converted cellulose and hydrolyze the substrate, their rate will be higher than compared to the uninterrupted experiment because their adsorption onto productive sites will be more favored. During hydrolysis all parts of the substrate (Figure 7) will change states; if there was an isolated and fixed hydrolysable portion, then we would have an upper cap on the maximum conversion that can be achieved, but that is clearly not the case. Due to this dynamic interchange of states between the accessible, hydrolysable, and non-accessible fractions of the substrate, the enzymes too will transition from one state to another and can get stuck/clogged in a dynamic fashion along conversion.

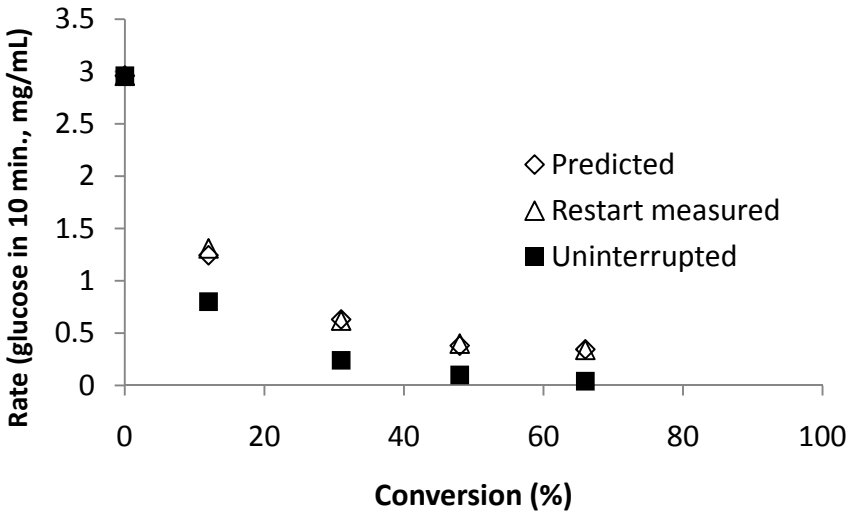


Figure 12. Uninterrupted rates, and predicted and measured restart rates at different conversion levels. Rate – glucose produced in 10 minutes, mg/mL. Uninterrupted rates were calculated by fitting an empirical curve to the hydrolysis curve (see Materials and Methods).

3.4. Summary of changes in accessibility, hydrolysability and reactivity

To summarize the three major factors investigated in this work, normalized parameter values of $[E]_{ads,max}$ (maximum adsorbable capacity), K_{ad} (adsorption equilibrium constant in the Langmuir isotherm), k (reactivity), and α (hydrolysability) as a function of conversion are shown in Figure 13. The fitted values of $[E]_{ads,max}$ did not vary strongly, implying only slight changes in the total accessible fraction of cellulose. The adsorption equilibrium constant K_{ad}, however, showed a marked decline with conversion, possibly indicating either a decrease in cellulase affinity for cellulose or an increase in the dissociation constant for cellulase-cellulose binding.

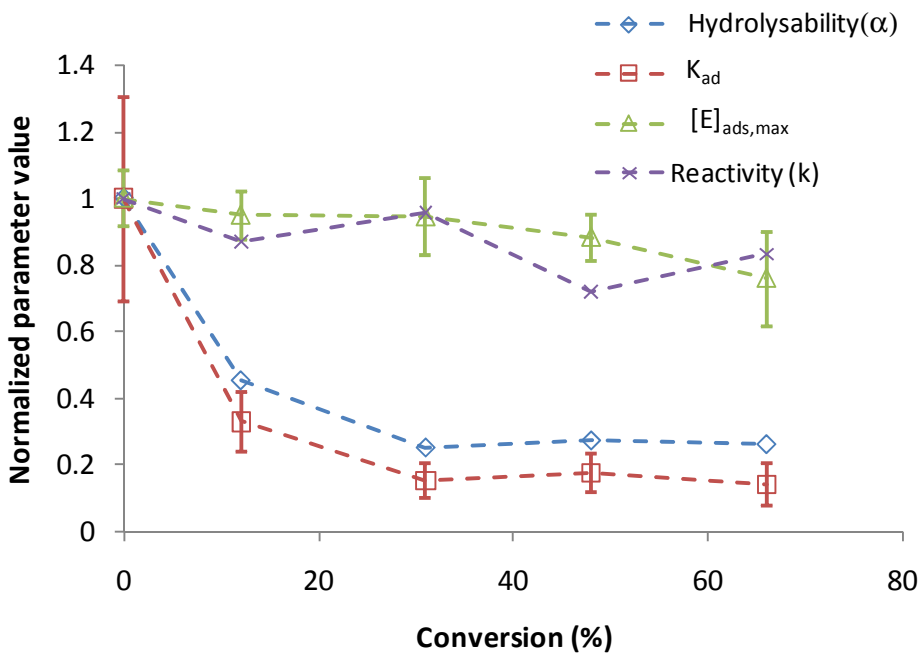


Figure 13. Normalized parameter values as a function of conversion.

Reactivity (k) did not show any noticeable trend with conversion. One of the major properties governing the rates, and the free energies associated with the hydrolysis steps (bond cleavage, decrystallization of chains for further hydrolysis, product expulsion (Beckham et al., 2011)), is crystallinity (Hall et al., 2010a). It is possible that the constancy of crystallinity and reactivity are strongly linked. Since there are other properties such as particle size, degree of polymerization, pore structures, microfibril morphology, etc. that can affect hydrolysis rates too (Mansfield et al., 1999), the relation between crystallinity and reactivity is not conclusive.

Hydrolysability (α) declined sharply from 100% to 25% until a degree of conversion of cellulose of 30%, beyond which it remained constant. For high enzyme loadings, and consequently high substrate coverage, all the productive sites will be exhausted and rates are expected to be governed by hydrolysability. An interesting feature of the parameters is that there is not a strong change beyond a conversion level of

approximately 30%. One could argue that considering this observation, the kinetic studies pursued in this thesis do not explain the rate decline beyond 30%. This interpretation, however, is incorrect as most of the rate decline is clearly accounted for by the restart experiments (Figure 12).

3.5. Prediction of rates using the developed kinetic rate law and role of clogging

If equation (3) is simplified for high enzyme loadings such that $f = \alpha/y$ (equation 8), then a simple expression relating the rates to reactivity k, accessibility $[E]_{ads,max}$ and hydrolysability α is obtained (equation 10). Since parameters change with conversion, they have to be determined with independent experiments at various conversion levels. This, along with the highly non-linear trend of the parameters with conversion, limits the predictive capability of the kinetic rate law in equation (3). It must be emphasized here that to start with, the aim of this chapter was to tease apart the various causes for rate slowdown, and not necessarily develop a universal expression predicting rate.

$$\text{Rate} = dX/dt = k*S_0(1-X)*[E]_{ads}*f \quad (3)$$

Substituting $f = \alpha/y = \alpha/([E]_{ads}/[E]_{ads,max})$, we obtain -:

$$\text{Rate} = dX/dt = k*[E]_{ads,max}*\alpha*S_0*(1-X) \quad (10)$$

One property of the kinetic studies pursued in this thesis, and hence of the parameters determined, is that they are based on restart experiments which can give rates different than uninterrupted ones on partially converted cellulose (Figure 12). As mentioned before, this could be due to the clogging phenomenon. To achieve close to 100% prediction of rates with equations 3&10, a model explaining the clogging phenomenon is needed.

Kurasin and Valjamae (2011) proposed that in the presence of obstacles, k_{off} (dissociation rate constant of cellobiohydrolases) governs the clogging phenomenon and

the hydrolysis rates, as the clogged enzymes need to desorb before they can bind productively to the substrate again. This simple scheme is likely to hold only for very low enzyme loadings. At high enzyme loadings, if the productive sites are completely covered, any desorbing enzyme is very likely to re-adsorb on to a non-hydrolysable (or clogged sites) portion of the substrate. Modeling clogging as a first order process is also unlikely to explain the phenomenon (Figure 4 b & 2). To incorporate clogging into a predictive model, further experimental investigation is needed into how cellulases get blocked or clogged on the cellulose surface.

3.6. Conclusions

Kinetic studies based on findings from modeling and experimental results provided unequivocal evidence to reveal accessibility and hydrolysability, and not reactivity, to be the major rate hindrances in the enzymatic hydrolysis of cellulose of Avicel. Hydrolysability, calculated as the fraction of the accessible cellulose that is reactive (resulting in productive adsorption), decreases from nearly 100% to approximately 25% at about 30% conversion. Reactivity, measured in terms of substrate hydrolyzed per amount of actively adsorbed cellulase, remains constant over the course of conversion. This is in congruence with our previous finding of constant crystallinity over the course of hydrolysis. While accessibility has been known to change with conversion and affect rates, hydrolysability has never been quantified. Clogging, or enzyme jamming at blocking sites on the substrate surface, is also an important phenomenon that could be experimentally verified through the faster rates obtained after restart compared to the uninterrupted rates. As the enzyme system employed was a cellulase mixture, differentiating between chain ends and bulk cellulose as reactive or nonreactive was not possible. While trends in accessibility and restart rates can vary between lignocellulosic substrates depending on the pretreatment method (Kumar and Wyman, 2009), the

methodology presented in this thesis (with a pure cellulosic substrate) can still be extended to lignocellulosic substrates.

Findings of the work presented here can also guide process strategies for cellulose biohydrolysis. Since hydrolysability and rates decrease by an order of magnitude for the first 30% cellulose converted, rendering the substrate less recalcitrant through biological pretreatment for the remaining 70% might be a very promising route (Hall et al., 2011). Engineering cellulases that can operate at higher temperatures (Hall et al., in press; Heinzelman et al., 2009) will result in an overall higher productivity (activity per amount of cellulose over the lifetime of the enzyme) and should be part of a strategy to render commercial process more economical.

CHAPTER 4

MULTIVARIATE STATISTICAL ANALYSIS OF X-RAY DATA FROM CELLULOSE: A NEW METHOD TO DETERMINE DEGREE OF CRYSTALLINITY AND PREDICT HYDROLYSIS RATES

(Experimental work associated with this chapter and Bansal et al. (2010) was carried out by Dr. Mélanie Hall)

4.1 Introduction

Crystallinity of cellulose is one of the major substrate properties governing the enzymatic hydrolysis rates and has been the focus of many works (Hall et al., 2010a; Lynd et al., 2002; Mansfield et al., 1999; Zhang and Lynd, 2004). Accurate quantification of the crystalline content in cellulose, termed crystallinity, is thus of prime importance, as it gives an estimation of the recalcitrance of the substrate to the enzymatic attack. Crystalline regions and lattices are formed due to hydrogen bonds between the cellulose chains and van der Waals forces between the glucose molecules. The degree of crystallinity, an average property, is the fraction of the crystalline content in the sample under consideration.

The techniques used for determining the degree of crystallinity of cellulose include X-ray powder diffraction, solid state ^{13}C-NMR, density measurements (Krassig, 1993) and more recently FT Raman spectroscopy (Schenzel et al., 2005), with X-ray diffraction being most widely followed. While ^{13}C-NMR is a reliable method for calculating crystallinity, it usually requires extensive acquisition time to obtain good peaks resolution and tends to be not applicable to low degrees of crystallinity, as the crystalline and amorphous peaks are hardly distinguishable. A number of methods to calculate the degree of crystallinity of cellulose from X-Ray diffraction spectra have been

published (Table 3). One major feature of all the methods (except for the peak height method (Segal et al., 1959) and method 1 of Wakelin et al. (1959)) is subtraction of the amorphous spectrum as background. While doing so by scaling the acquired spectrum of an amorphous polymer (to bring it below the crystalline spectrum) may be physically meaningful for spectra with sharp peaks, for cellulose it is not as simple due to considerable peak overlaps (the different crystal planes for cellulose I are labeled in Figure 14). With the advent of software programs such as JADE®, functional deconvolution of spectra with respect to a chosen background is simple; the issue then is the choice of the background. The easy-to-use method of Segal et al. (1959), which is still the most widely used, does not need background subtraction but the definition of a baseline, and is based on peak heights. The degree of crystallinity of cellulose I is given by comparing the minimum in intensity above baseline at $2\theta = 18^{o}$ (I_{am}), and the maximum in intensity at $2\theta = 22.5^{o}$ (I_{200}), accounting for the amorphous part and the crystalline part (major diffraction from the 200 plane) respectively ($CrI = 100*(I_{200}-I_{am})/I_{200}$). However, it is clear from Figure 14 that the trough at 18^{o}, which is assumed to account for the amorphous portion, is shifted to lower angles compared to the actual reflection from a pure amorphous sample (maximum intensity at $2\theta = 19.5°$). Nevertheless, the method is useful for relative comparison and results should be carefully interpreted when used for absolute crystallinity index determination.

Thygesen et al. (2005) applied four different methods (peak-height method (Segal et al., 1959), Ruland-Vonk (Ruland, 1961; Vonk, 1973), profile refinement method (Rietveld, 1969) and Debye calculations (Debye, 1915)) to calculate the crystallinity index of four different substrates and showed that the results can vary within a range of up to 20% depending on the method. For Avicel, a pure microcrystalline cellulose I sample and one of the substrates used for the work in this chapter and in (Bansal et al., 2010), calculated crystallinities from X-ray spectra in the literature vary over a large range (54% – 92.97%) (Table 4). Although differences in the ways to handle the samples

(drying method and temperature, amount of sample analyzed) are likely to introduce variations in the results, the work of Thygesen et al. (2005) shows that different crystallinity values can be extracted using different analytical methods on the same spectrum. This was also demonstrated recently by (Park et al., 2010).

Given the importance of cellulose crystallinity in the enzymatic hydrolysis and its role in evaluating efficient pretreatment methods (or how to render a cellulose sample more amorphous), the topic of crystallinity calculation from X-ray data has been revisited. We use X-ray powder diffraction for measuring the intensities of beams (averaged over the sample used in the setup) at various diffraction angles to calculate the degree of crystallinity (weight fraction of the crystalline content). While it may be possible to calculate microscopic properties such as the crystallite dimensions corresponding to different phases (Garvey et al., 2005) and structural determination in terms of atomic coordinates (Nishiyama et al., 2002) using the X-ray diffraction data, we do not try to quantify any microscopic property of cellulose other than its degree of crystallinity.

The data-driven method developed utilizes X-ray diffraction spectra of cellulose samples of intermediate crystallinity prepared by treating Avicel and fibrous cellulose (FC) with varying (and controlled) concentrations of phosphoric acid. Purely amorphous samples were obtained for both types of cellulose. To calculate crystallinity indices, normalized X-ray diffraction spectra were expressed as a linear combination of the normalized untreated cellulose (Avicel or FC) and amorphous cellulose spectra. Principal component analysis (PCA) was also applied to the spectroscopic data (separately to Avicel and FC spectra sets) and the principal component scores were related to calculated crystallinities. This revealed the dimensionality of the X-ray spectra data. Cellulose mixtures with varying fractions of untreated and amorphous Avicel were prepared to validate the prediction of crystallinity values. Based on the observation that initial hydrolysis rates followed a linear trend with the calculated crystallinity index, principal

component regression was used to successfully predict the initial hydrolysis rates from X-ray spectra.

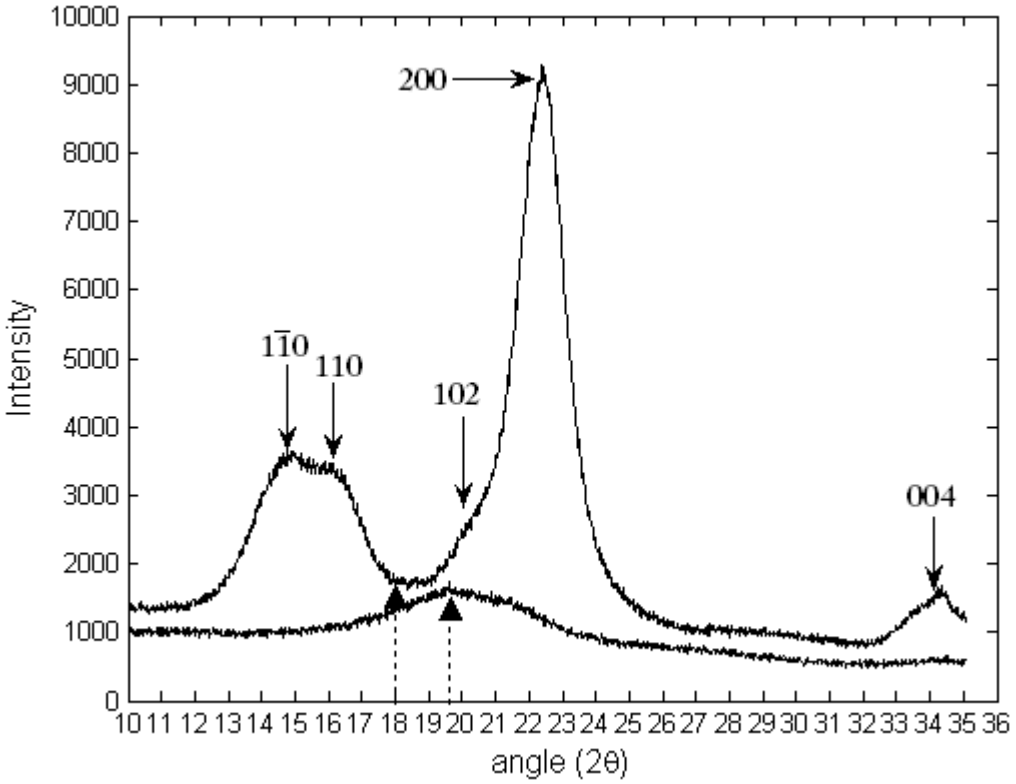


Figure 14. X-ray spectra of Avicel (upper spectrum) and amorphous cellulose (lower spectrum) with major crystal planes labeled with solid arrows. Dashed arrows show locations of intensity minimum in Avicel spectrum at 18° and intensity maximum in amorphous cellulose spectrum.

Table 3. Published methods on the calculation of crystallinity index from X-ray spectra (for detailed explanations, the reader may refer to the original works).

Reference	Mathematical methodology
Hermans and Weidinger (1948)[a]	Total crystalline and amorphous intensity is calculated from the diffraction pattern (area under the curve) by marking the crystalline and amorphous portions in the spectra. Crystallinity is then expressed as -: $X_c = \frac{I_c}{I_c + K*I_a}$ (1) where X_c – crystallinity, I_c – crystalline portion intensity, I_a – amorphous portion intensity, K – empirical constant
Segal et al. (1959)[a]	Ratio of intensities at $2\theta = 22.5^o$ to that at $2\theta = 18^o$ gives the ratio of crystalline to amorphous fractions (cellulose 1) or $2\theta = 19.5^o$ to that at $2\theta = 16^o$ (cellulose II). 2θ – diffraction angle: $CrI = 100*(I_{200}-I_{am})/I_{200}$ (2) I_{am} - minimum in intensity above baseline at $2\theta = 18^o$, I_{200} - maximum in intensity above baseline at $2\theta = 22.5^o$ (I_{200}),
Wakelin et al. (1959)[a]	Method 1: Correlation of the difference in the intensities of sample and amorphous with difference in intensities of crystalline standard and amorphous. Method 2: Area between the sample spectrum and the amorphous spectrum. Relative crystallinity is given by the ratio of this area to that calculated with the crystalline standard
(Ruland, 1961; Vonk, 1973)	Separation of amorphous and crystalline spectra. Amorphous spectrum scaled to match the spectrum of partially crystalline sample at regions where peaks are absent. Crystallinity is given by -: $X_c = \frac{\int_{s_o}^{s_1} s^2 I_c ds}{\int_{s_o}^{s_1} s^2 I\, ds} \times \frac{\int_{s_o}^{s_1} s^2 \overline{f}^2 ds}{\int_{s_o}^{s_1} s^2 \overline{f}^2 D^2 ds}$ (3) where X_c – crystallinity, I_c – Intensity of crystalline portion, I – Total intensity, $s = 2\sin\theta/\lambda$, 2θ – diffraction angle, $\overline{f}^2$ - mean square of scattering, D – disorder function
Chung and Scott (1973)	Amorphous spectrum expressed as a Gaussian like function and is subtracted as background from the sample spectrum. Crystallinity and a constant k determined by use of following equations: $I_a = k_a x_a$ (4) $I_c = k_c x_c$ (5) $x_c + x_a = 1$ (6) I_a - amorphous portion intensity (area under the diffraction curve), I_c – crystalline portion intensity, x_a – fraction of amorphous component, x_c – fraction of crystalline component, k_a, k_c- constants.

Table 3 (continued)

Reference	Mathematical methodology
Soltys et al. (1984)[a]	Crystalline diffraction pattern was obtained after removing the linear background and scaling the amorphous sample spectrum. Crystallinity calculated as the ratio of area under the crystalline diffraction peaks to the total area.
Polizzi et al. (1990)	The background is expressed as a function of the amorphous spectra, and has crystallinity and disorder factor as parameters. These two parameters along with the parameters of the fitting function for the sample spectrum are optimized for the best fit. The background is given by -: $I_B(s) = (1-X_c)I_{am}(s) + X_c \langle f(s^2) \rangle \{1 - \exp(-ks^2)\}$ (7) $I_B(s)$ – background scattering, X_c – degree of crystallinity, $I_{am}(s)$ – experimental intensity of amorphous sample, $\langle f(s^2) \rangle$ - mean square atomic scattering factor, k – disorder factor, $s = 2\sin\theta/\lambda$, 2θ – diffraction angle.
Majdanac et al. (1991)[a]	Instrument background subtracted from spectrum, amorphous scattering expressed by a Gaussian function, the peaks in the spectra expressed as Gaussian or Lorentzian functions. Crystallinity is given by the area under the curves (other than the amorphous Gaussian curve) divided by the total area

[a]Developed for cellulose

Table 4. Crystallinity values of Avicel in literature calculated from X-ray spectra (works reporting relative crystallinity values are not tabulated).

Reported crystallinity (%)[b]	**Reference**	**Method used**
81.3 (PH-101)	Gama et al. (1994)	Segal et al. (1959)
77.7 (PH-101) 80.1 (PH-105)	Schurz and Klapp (1976)	Area of the amorphous background is subtracted from the X-ray diffraction curve by drawing the background curve such that it joins the points where peaks are absent.
77 (PH-101) 75.3 (PH-102) 73.8 (PH-103) 72.8 (PH-105)	Soltys et al. (1984)	Soltys et al. (1984)
82 (PH-101)	Doelker et al. (1987)	Hermans and Weidinger (1948)
54 (PH-101)	Teeäär et al. (1987)	Ruland (1961)
92.97 (PH-101)	Dourado et al. (1998)	Segal et al. (1959)
81 (PH-101)	(Marson and Seoud, 1999)	Segal et al. (1959)
72.23 (PH-101)	Kumar et al. (2001)	Area under the peaks of crystalline reflections expressed as a percentage of the area under the hydrocellulose curve (taken to be 100% crystalline reference).
83 (PH-101)	Ramos et al. (2005)	Segal et al. (1959)
82 (PH-101)	Gupta and Lee (2008)	Segal et al. (1959)
62	Thygesen et al. (2005)	Segal et al. (1959)
67	Thygesen et al. (2005)	(Ruland, 1961; Vonk, 1973)
41	Thygesen et al. (2005)	Rietveld (1969)
39	Thygesen et al. (2005)	Debye (1915)
64 (PH-102)	Ardizzone et al. (1999)	-
63 (PH-102)	Nakai et al. (1977)	Hermans and Weidinger (1948)
87.6	Souza et al. (2002)	Segal et al. (1959)

[b] The type of Avicel used is shown in parenthesis for the works which report it

4.2 Materials and methods

Information on chemicals and materials, phosphoric acid pretreatment, enzymatic hydrolysis of cellulose, determination of glucose content, X-ray diffraction, solid state ^{13}C NMR can be found in Appendix C. Data analysis and calculations are described in this section.

4.2.1 Data normalization

The data - intensities of the X-ray spectra at different diffraction angles, were normalized with respect to the area under the intensity-angle curve. Since the control over the small amount of sample put on the X-ray diffractometer is not easy to achieve, the intensity for any given sample can vary. The area was calculated by the following expression:

$$\text{Area} = h * \sum_{i=1}^{n} I(2\theta_i) \tag{8}$$

where h is the scanning step size of the diffraction angle (0.0167 ^{o}), $I(2\theta_i)$ is the intensity at the diffraction angle $2\theta_i$, n is the number of scan points (2992 for $2\theta = 10^o$ to 60^o and 1500 for $2\theta = 10^o$ to 35^o; as will be explained in section 4.3.2, we ultimately use data only up until 35^o for our analysis).

4.2.2 Calculation of the crystallinity index

The normalized spectrum of a sample was expressed as a linear combination of the normalized spectra of commercial cellulose (Avicel or FC) and amorphous cellulose (phosphoric acid swollen cellulose (PASC)) samples:

$$I_j(2\theta) = f_j \, I_P(2\theta) + (1 - f_j) I_C(2\theta) + \varepsilon \tag{9}$$

where $I_j(2\theta)$ is intensity of the j^{th} sample at diffraction angle 2θ, $I_p(2\theta)$ is intensity of PASC at diffraction angle 2θ, $I_C(2\theta)$ is intensity of pure cellulose at diffraction angle 2θ, f_j is contribution of PASC to the spectrum, ε is random error.

$\hat{f}_j$, the least square estimate of f_j, was used to estimate the crystallinity by multiplying the contribution of Avicel or FC $(1-\hat{f}_j)$, to their crystallinity (taken to be 60% for Avicel and 72% for FC - as measured by ^{13}C-NMR):

$$\mathrm{Cri_j} = (1-\hat{f}_j) * \mathrm{Cri_C} \tag{10}$$

where $\mathrm{Cri_j}$ is crystallinity (in percentage) of the j^{th} sample of Avicel or FC, $\mathrm{Cri_C}$ is the crystallinity of Avicel or FC depending on the cellulose under consideration.

The numerical estimation of f_j is similar to the estimation of the slope in the correlation method of Wakelin et al. (1959). Crystallinity was also calculated from the reconstructed spectra after performing the principal component analysis. For comparison, the peak-height method (Segal et al., 1959) was also applied:

$$\mathrm{Cri} = 100*(I_{200}-I_{am})/I_{200} \tag{2}$$

where I_{200} is the maximum intensity above baseline at $2\theta = 22.5^o$ and I_{am} is the minimum in intensity above baseline corresponding to amorphous content at $2\theta = 18^o$.

4.2.3 Principal component analysis of X-ray spectra

The X-ray spectra data (normalized), expressed as given below, was subject to principal component analysis. More details of PCA can be found elsewhere (Jolliffe, 2002).

$$X^T = \begin{array}{c} \begin{matrix} 2\theta_1 & 2\theta_2 & & 2\theta_n \end{matrix} \\ \begin{pmatrix} x_{11} & x_{12} & \ldots\ldots\ldots & x_{1n} \\ \cdot & & & \\ \cdot & & & \\ x_{m1} & x_{m2} & \ldots\ldots\ldots & x_{mn} \end{pmatrix}_{m\times n} \end{array} \tag{11}$$

where m is the number of samples (17) for Avicel and (15) for FC, n is the number of scan points (1500).

First, the data matrix is transformed into the following Z matrix by mean centering the columns and making them of unit variance:

$$Z^T = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{s_1} & \frac{x_{12}-\bar{x}_2}{s_2} & \cdots\cdots & \frac{x_{1n}-\bar{x}_n}{s_n} \\ . & & & \\ . & & & \\ \frac{x_{m1}-\bar{x}_1}{s_1} & \frac{x_{m2}-\bar{x}_2}{s_2} & \cdots\cdots & \frac{x_{mn}-\bar{x}_n}{s_n} \end{pmatrix}_{m\times n} \tag{12}$$

$\bar{x}_1$ and s_1 are the mean and standard deviation of the first column in X^T and so on. To find the principal components (vectors in the n dimensional space along which the data might be concentrated), singular value decomposition (SVD) of the Z matrix (given below) is performed:

$$Z=U_{n\times n}S_{n\times m}V^T_{m\times m} \tag{13}$$

U and V are matrices consisting of orthonormal vectors.

$$S = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & 0 & & 0 \\ . & & . & & & \\ . & & & . & & \\ 0 & & & & \sigma_m & \\ 0 & & & & 0 & \\ \vdots & & & & \vdots & \\ \vdots & & & & \vdots & \\ 0 & & & & 0 & \end{pmatrix}_{n\times m} \tag{14}$$

$\sigma's$ are the singular values.

The column vectors in U corresponding to the relatively larger values of $\sigma's$ represent the principal component directions.

The score matrix is calculated by projecting the columns of Z onto the directions of the principal components as below:

$$Z_{tf} = U^{T}_{L} Z = S_L V^{T}_{L} \quad (15)$$

where L denotes that we have taken only the first L principal $\sigma's$.

As will be explained later, for computing the Z matrix, the columns were not normalized with the standard deviation. The other steps are the same as above.

4.2.4 Calculation of crystallinity index from principal components

The Z and X matrix can be reconstructed from the reduced dimensional score matrix Z_{tf}:

$$Z_r = U_L Z_{tf} = U_L U^{T}_{L} Z \quad (16)$$

$$(X_r)_{ij} = (Z_r)_{ij} + \overline{x}_i \quad (17)$$

Using the reconstructed X data, equations (9) and (10) were applied to calculate the crystallinity. This is equivalent to regression of the crystallinity to the component scores in equation (9) instead of the whole spectra.

4.2.5 Principal component regression (PCR) for predicting hydrolysis rates

Prediction of hydrolysis rates by regressing them to X-ray data is not possible since the number of equations (17 for Avicel and 15 for FC, corresponding to the number of available samples) will be far less than the number of parameters (1500 for intensity at each diffraction angle plus one constant). However if the data is projected onto the first few principal components, hydrolysis rates can be regressed to the scores of the different samples in the direction of the principal components:

$$h_j = \beta_o + \beta_1 C_{j1} + \beta_2 C_{j2} + \ldots + \beta_L C_{jL} + \varepsilon_j \quad (18)$$

j = 1,2….,m, h_j is the initial hydrolysis rate of the j^{th} sample, β's are the regression parameters, C_{ji} is the score of the j^{th} sample's X-ray spectra in i^{th} principal component's direction, and ε_j is random error.

4.2.6 Principal component analysis and principal component regression on the combined spectra sets of Avicel and FC

PCA and PCR were performed on the combined spectra sets of Avicel and FC in order to have a mathematical framework applicable to both the cellulose I substrates. As will be shown in section 3.6, only two PCs were used after PCA. For calculation of crystallinity index, the reconstructed spectra from two PCs were used in the following equations:

$$I_j(2\theta)_r = f_{j1}I_P(2\theta)_r + f_{j2}I_A(2\theta)_r + f_{j3}I_{FC}(2\theta)_r + \varepsilon \quad (19)$$

$$f_{j1} + f_{j2} + f_{j3} = 1 \quad (20)$$

where $I_j(2\theta)_r$ is intensity of the j^{th} sample at diffraction angle 2θ, $I_p(2\theta)$ is intensity of amorphous cellulose at diffraction angle 2θ, $I_A(2\theta)$ is intensity of Avicel at diffraction angle 2θ, r denotes that the spectrum is reconstructed, f_{j1}, f_{j2} *and* f_{j3} are the contributions of amorphous cellulose, Avicel and fibrous cellulose respectively to the spectrum of the j^{th} sample, ε is random error.

The least square estimates of f_{j2} and f_{j3} - $\hat{f}_{j2}$ and $\hat{f}_{j3}$ were then used to calculate the crystallinity of the j^{th} sample:

$$Cri_j = \hat{f}_{j2} * Cri_A + \hat{f}_{j3} * Cri_{FC} \quad (21)$$

where Cri_j is the crystallinity index of the j^{th} sample, Cri_A (=60%) and Cri_{FC} (=72%) are the reference crystallinity indices of Avicel and FC respectively.

Hydrolysis rates calculation was performed by equation 18 where the scores used were the ones obtained by PCA of the combined spectra sets of Avicel and FC.

All the data analysis and calculations were done in MATLAB® (The Mathworks Inc. R2008b).

4.3. Results and discussion

4.3.1 Phosphoric acid pretreatment of cellulose samples

To obtain a reliable analytical method to accurately determine cellulose degree of crystallinity, several intermediate crystallinity indexes of a same cellulose source are required so that the method can be tested and validated over the whole range of crystallinity indices. Observed variations in X-ray diffraction are then attributed exclusively to the variation in crystallinity. It has been shown in other works that phosphoric acid pretreatment does not alter the degree of polymerization (and thus the molecular weight) of cellulose (Jeoh et al., 2007; Zhang and Lynd, 2005). This property can therefore be assumed to be constant within all the acid pretreated samples. Pretreatment of Avicel and FC with increasing concentrations of phosphoric acid resulted in subsequent decrease of the degree of crystallinity, as observed by a significant change in the X-ray diffraction spectra (Figure 15) and applying the peak height method to the X-ray diffraction spectra (Table 5, columns 8 and 9). The enzymatic hydrolysis rates obtained with cellulases (supplemented with an excess of β-glucosidase) increased consistently with the acid concentration used for pretreatment of the corresponding samples (Figure 16), suggestive of a relationship between hydrolysis rates and initial degree of crystallinity. A sigmoidal type of behavior of hydrolysis rates with respect to phosphoric acid concentrations (i.e. pretreatment conditions) was observed, where major changes in substrates seemed to occur over a very narrow range of phosphoric acid concentrations (75% to 82 %). The amorphous references (0% crystalline PASC for crystallinity index calculations) for Avicel and FC were obtained from pure Avicel treated with 82.37 % phosphoric acid (Avi2) and pure FC treated with 85.00% phosphoric acid (FC1) respectively. For Avicel, increasing the phosphoric acid

concentration beyond that point did not lead to any additional changes in the spectrum. For FC, 82.03% phosphoric acid also led to an amorphous sample. Both amorphous cellulose spectra could be superimposed (Figure 31, Appendix D). The X-ray diffraction pattern of amorphous cellulose (phase I) seems therefore to be independent from the original cellulose from which it is obtained by acid pretreatment. Also, cellulase reactivity was identical on amorphous Avicel (Avi2) and FC (FC2) (9.23 and 9.16 mg/ml resp., Table 5). Although the amorphous sample (0%, FC1) obtained from fibrous cellulose (85% phosphoric acid) had an unexpected lower hydrolysis rate of 7 mg/ml, no considerable change in reactivity (hydrolysis rates) beyond phosphoric acid concentration of 81.71% was observed (rates were 9.74, 9.36 and 9. 16 mg/ml resp. for FC treated with 81.71%, 81.78% and 82.03% phosphoric acid resp., Table 5 b column 3). Peak height method gave a degree of crystallinity of 0% and 21% for amorphous cellulose obtained from Avicel (Avi2) and FC (FC1) respectively. Although the peak height method does not give a value of 0% for FC1 and gives the lowest value for FC4 (6%), FC1 was taken as the amorphous reference since the spectrum could be superimposed on amorphous Avicel much better than FC4, did not present any significant peaks and had the highest acid concentration (FC4 presented a small shoulder at $2\theta = 23°$, Figure 31 in Appendix D). Moreover, there can be variations in the calculations from the peak height method for samples with lower degrees of crystallinity since there is no detectable trough/minimum near the diffraction angle of 18^{o}, and a shift by less than one degree for I_{am} in equation 11 can cause significant changes in the calculated crystallinity; this is shown in Table 5 (columns 8 and 9) for values calculated independently by two of the authors, where great variations were seen in the calculated crystallinity index for samples treated with high phosphoric acid concentrations (close to being amorphous). The baseline selection is also error-prone.

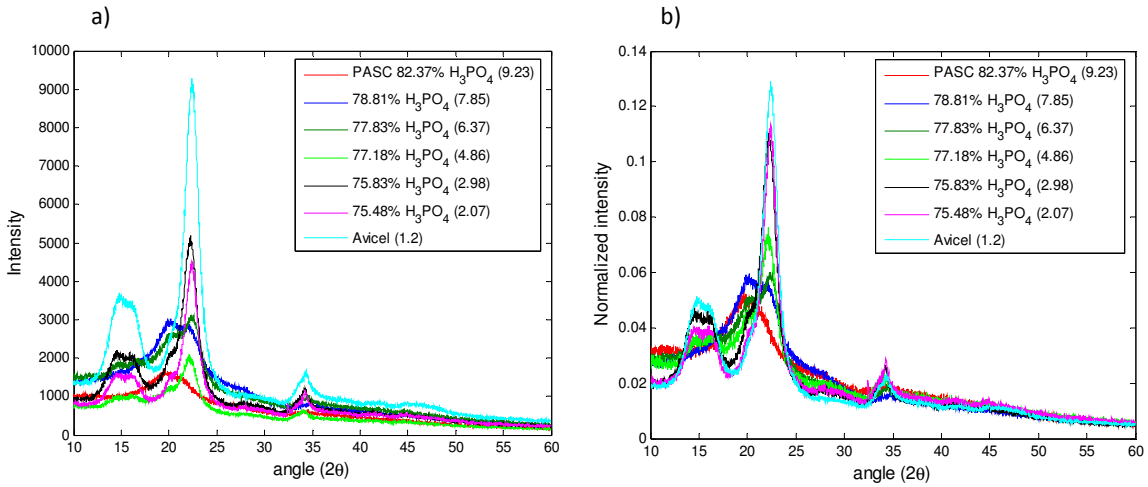


Figure 15. X-ray spectra of various phosphoric acid pretreated Avicel samples a) before and b) after normalization. Hydrolysis rates are shown in parenthesis (mg/ml of glucose produced in the first 2 minutes of the reaction with cellulases).

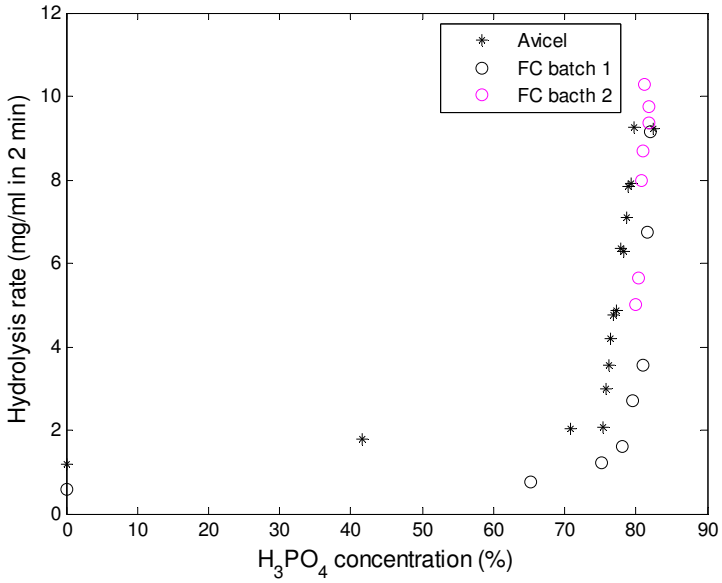


Figure 16. Hydrolysis rates of phosphoric acid-pretreated Avicel and FC vs. H_3PO_4 concentrations used for pretreatment. Two different commercial phosphoric acid solutions (85% w/w) were used for FC and are shown in different colors. Note: two samples of FC obtained from these two undiluted solutions gave unexpectedly lower hydrolysis rates of around 7 mg/ml glucose in 2 min.; these points were not incorporated in the analysis. The samples were still shown to be amorphous (X-ray data) and one of them was taken as the reference amorphous cellulose.

Table 5. Crystallinity values obtained from various methods and corresponding hydrolysis rates of various phosphoric acid pretreated samples for a) Avicel and b) FC. Columns correspond to: 1 – Sample name, 2 (Acid) – Acid concentration (%), 3 (Rate) – Hydrolysis rate (mg/ml of glucose produced in the first 2 minutes), 4 – Cri (PCA) (%), 5 – Cri (All data) (%), 6 – Cri (LOO) (%), 7 – Cri (Avicel subtraction) (%), 8 – Cri (Segal method) (%) 1^c, 9 – Cri (Segal method) (%) 2^c.

a)

1	2 (Acid)	3 (Rate)	4	5	6	7	8^c	9^c
Avi1	82.41	8.74	5.28	5.98	5.40	5.10	19.57	47.37
Avi2 (PASC)	82.37	9.23	0.00	0.00	0.00	0.00	0.00	0.00
Avi3	79.64	9.27	1.12	1.08	1.18	0.79	16.28	10.00
Avi4	79.23	7.9	6.53	7.45	6.71	6.46	29.31	32.5
Avi5	78.81	7.85	7.97	9.08	8.13	8.15	35.71	49.37
Avi6	78.6	7.1	10.89	11.95	11.11	10.83	40.43	45.28
Avi7	78.35	6.3	16.45	17.61	16.57	16.96	54.42	50
Avi8	77.83	6.37	14.95	15.31	14.98	14.95	59.46	63.41
Avi9	77.18	4.86	22.97	23.18	22.99	22.98	78.21	78.84
Avi10	76.79	4.76	20.41	20.65	20.53	20.63	72.84	79.59
Avi11	76.49	4.2	32.66	31.79	32.61	31.57	85.28	83.33
Avi12	76.12	3.55	41.95	42.02	41.88	42.05	85.39	85.45
Avi13	75.83	2.98	44.94	45.67	44.64	45.28	88.76	88.89
Avi14	75.48	2.07	49.97	49.41	49.79	49.18	93.33	89.10
Avi15	70.81	2.05	54.43	52.75	53.11	52.52	91.26	90.32
Avi16	41.56	1.79	56.52	54.89	55.80	54.60	93.58	90.16
Avicel	0	1.2	60.00	60.00	60.00	60.00	92.31	90.98

Table 5 (b)

1	2 (Acid)	3 (Rate)	4	5	6	7	8[c]	9[c]
FC2	82.03	9.16[a]	-0.48[b]	-0.33	-0.41	-0.59	22.99	10.34
FC3	81.78	9.36	1.71	1.85	1.76	1.36	32.74	26.32
FC4	81.71	9.74	-0.86[b]	-0.62	-0.66	-1.38	5.94	5.88
FC5	81.5	6.74	11.42	11.26	11.57	10.47	57.32	55.36
FC6	81.22	10.29	7.33	7.26	7.34	7.00	58.57	55.55
FC7	81.06	8.7	12.63	12.69	12.63	12.42	72.97	73.58
FC8	81.05	3.55	49.55	48.35	49.49	48.13	91.98	90.48
FC9	80.71	7.97	25.83	25.66	25.85	25.51	81.18	80.00
FC10	80.46	5.63	36.01	35.95	36.01	35.46	85.96	84.38
FC11	79.94	5.02	40.32	40.41	40.31	39.98	86.89	84.61
FC12	79.51	2.72	64.52	62.99	64.37	62.49	94.24	93.06
FC13	78.09	1.61	68.68	67.44	68.64	67.16	95.83	95.45
FC14	75.07	1.22	72.46	71.01	72.15	70.19	95.41	94.53
FC15	65.22	0.77	74.03	72.08	73.65	71.39	95.88	95.31
FC	0	0.57	72.00	72.00	72.00	72.00	96.94	96.25

[a] A lower hydrolysis rate was obtained from the 85% phosphoric acid pretreated fibrous cellulose sample(7 mg/ml). This may be due to higher viscosity of the acid solution preventing a thorough wash of the cellulose sample (remaining acid would most likely impact the enzymatic activity).
[b] The 0% reference was obtained from 85% phosphoric acid pretreated fibrous cellulose and negative values are attributed to analytical error (< 1%)
[c] The Segal method (peak height method) was independently applied by two of the authors (1 and 2)

4.3.2 X-ray data normalization and calculation of crystallinity index

Figure 15 shows some of the X-ray spectra collected with samples of Avicel before and after normalization. Before normalization, though it is possible to differentiate qualitatively between the samples, a quantitative comparison is not possible, mainly due to the fact that overall intensity and initial intensity differ from one spectrum to another. After normalization, however, the spectra were brought to the same scale, making direct comparison possible. Moreover, relative to untreated Avicel, the normalized intensities clearly increased with acid concentration (and hydrolysis rates) in some intervals of the diffraction angle and decreased in others. Over most of the diffraction angle range, the intensities changed monotonically between untreated Avicel and amorphous cellulose at

any given diffraction angle (similar observations were made for FC). As hydrolysis rates were shown to be strictly related to phosphoric acid concentrations (Figure 16), their relationship with degrees of crystallinity was closely looked at. The correlation between the hydrolysis rates and the intensity values at various X-ray diffraction angles shows that the spectra intensities are highly correlated with the hydrolysis rates (Figure 17). The correlations are positive for the regions where hydrolysis rates increase with intensities (associated with a decrease in cellulose crystallinity), and negative for the regions where the hydrolysis rates decrease with intensities (peaks - associated with an increase in cellulose crystallinity). Although overall cellulose crystallinity arises due to contributions from different crystal planes (Figure 14) associated with different diffraction angle intervals in the X-ray spectra (Schurz et al., 1987), no attempt was made to attribute a physical meaning to the intervals in which the increase or decrease of intensity was observed. While some of these intervals that contain peaks could correspond to the major planes, the others could be just the tails of the peaks.

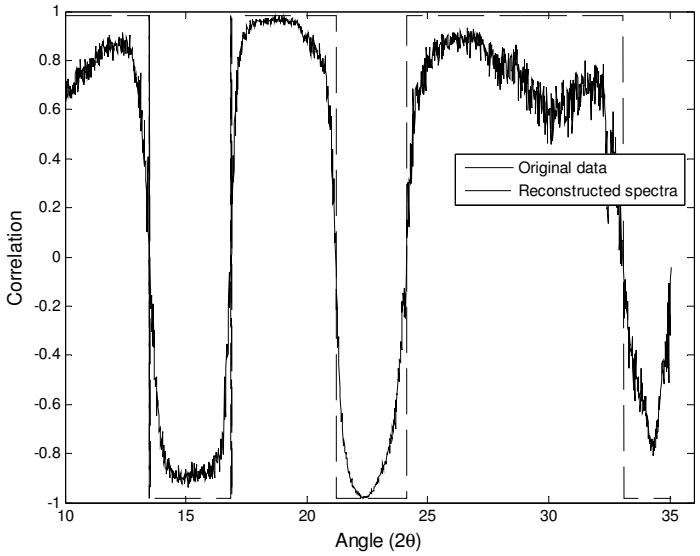


Figure 17. Correlation of the hydrolysis rates with intensities at different diffraction angles for the original spectra and the spectra reconstructed from PCA for Avicel. For FC the reconstructed curve was within the limits of -0.97 and 0.97 (see Figure 32 in Appendix D).

Beyond 35^{o} no considerable correlation was observed, therefore, only the data up to 35^{o} was considered for subsequent analysis. One of the implications of Figure 17 is that the spectrum of an intermediate crystalline cellulose sample can be represented as a linear combination of the spectra of commercial and amorphous cellulose:

$$\mathrm{I_j}(2\theta) = f_j\ \mathrm{I_P}(2\theta) + (1 - f_j)\mathrm{I_C}(2\theta) + \varepsilon \qquad (9)$$

The crystallinity is then calculated by:

$$\mathrm{Cri_j} = (1 - \hat{f}_j) * \mathrm{Cri_C} \qquad (10)$$

To have an X-ray independent value for the crystallinity index of untreated cellulose, reference crystallinity indexes ($\mathrm{Cri_C}$) for Avicel and FC were calculated from ^{13}C-NMR and were 60% and 72% respectively. Table 5 (column 5) shows the degree of crystallinity as calculated by equation (10) for Avicel and FC, and Figure 18 shows the gradual decrease in crystallinity index obtained from that method with the respective

increase in hydrolysis rates. The two sets of data were highly correlated as a linear fit with high R^2 value was obtained (> 0.95) (the 95% confidence intervals for slopes were -7.26 ± 0.78 and -7.88 ± 1.02 for Avicel and FC respectively, for the intercept the 95% confidence intervals were 64.95 ± 4.61 and 78.80 ± 6.64 for Avicel and FC respectively; p-value for the full model was of the order of 10^{-10} for both Avicel and FC, thus the linear model is statistically significant). For comparison, the peak height method values are also shown in Table 5 (column 8 and 9). Impractically high values for the degree of crystallinity were obtained from the peak height method, leaving the samples with higher values hardly distinguishable. For example, the last four Avicel samples in Table 5a were obtained from using different phosphoric acid pretreatment concentrations (Avi14 - 75.48%, Avi15 - 70.81 %, Avi16 - 41.56% and Avicel - 0%), and display accordingly decreasing hydrolysis rates (Avi14 – 2.07, Avi15 – 2.05, Avi16 – 1.79 and Avicel – 1.2 mg/ml), but the peak height method gave similar values for the four samples (Avi14 - 93.33%, Avi15 - 91.26%, Avi16 - 93.58% and Avicel - 92.31%, as calculated by 1); the method developed herein shows however a trend following the hydrolysis rates and a change of up to 10 percentage points in the degrees of crystallinity (Avi14 – 49.97%, Avi15 – 54.43%, Avi16 – 56.52% and Avicel – 60%). The R^2 values of the fits of spectra with equation (9) are overall high, indicating a good fit (≥ 0.9 for both Avicel and for FC, Table 14 in Appendix D). The fit of Avicel crystallinity with hydrolysis rates slightly differs from that of FC (y = -0.0726x + 0.6495 vs. y = -0.0788x + 0.788 resp.), implying that Avicel and FC have different enzymatic hydrolysis rates for the same crystallinity index (the amorphous cellulose samples however, display similar rates). Crystallinity is certainly not the only structural feature determining the enzymatic susceptibility of cellulose, and other factors like degree of polymerization (DP), accessibility to enzymes, and particle size can also play an important role (Mansfield et al., 1999; Zhang and Lynd, 2004).

The spectrum of a sample can be expressed as a linear combination of the spectra of Avicel/FC and amorphous cellulose, therefore, the amorphous background is composed of the pure amorphous cellulose spectrum plus the amorphous fraction in the commercial cellulose ($100 - Cri_C$). While most of the works cited in Table 3 also subtract a background to calculate the crystallinity indices, they do so by scaling the amorphous sample curve; in this work, contribution of the amorphous sample to the X-ray diffraction curve was very apparent by observing normalized spectra (Figure 15) and correlation curves (Figure 17), and no scaling was needed after normalization. The equivalence of equation (9) to the method of computing areas under the curves can be seen by integration with respect to the diffraction angle on both sides:

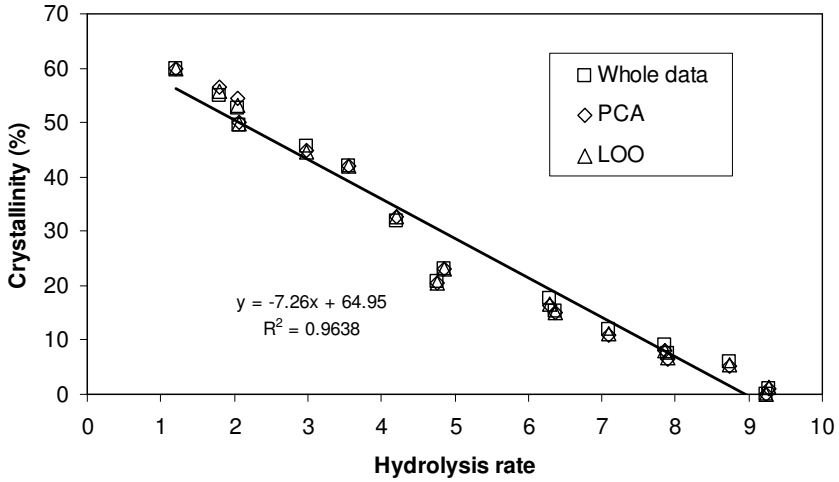
$$\int_{\theta_1}^{\theta_2} I_j(2\theta)d(2\theta) = f_j \int_{\theta_1}^{\theta_2} I_P(2\theta)d(2\theta) + (1-f_j)\int_{\theta_1}^{\theta_2} I_C(2\theta)d(2\theta) + \int_{\theta_1}^{\theta_2} \varepsilon d(2\theta) \quad (22)$$

The third term on the right hand side is equal to zero since the random error has a mean of zero. Equation (22) then can be written as:

$$A_j = f_j A_P + (1-f_j)A_C \quad (23)$$

where A_j is the area under the diffraction curve of the j^{th} sample, A_P is the area under the diffraction curve for PASC, A_C is the area under the diffraction curve for Avicel/ FC.

a)



b)

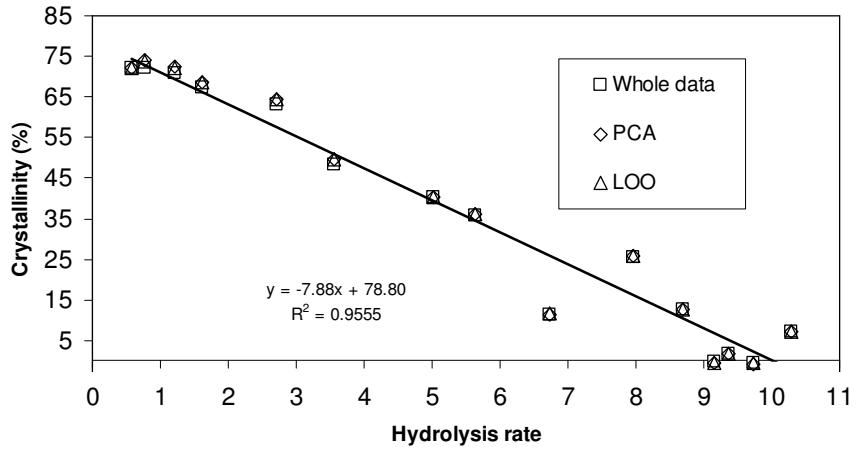


Figure 18. Calculated crystallinities vs. enzymatic hydrolysis rates: whole spectra in equations (9) and (10) (□), PCA (◊) and leave-one-out validation (LOO) (△) for a) Avicel and b) FC. (Hydrolysis rates correspond to the amount of glucose produced in the first 2 min of the reaction with cellulases). The linear equations shown are the fits between degrees of crystallinity calculated with whole spectra and hydrolysis rates.

4.3.3 Principal component analysis of X-ray data and calculation of crystallinity index from the principal component scores

Principal component analysis (PCA) reduces the dimensionality of data of which variables are correlated, by transforming the data set to a new orthogonal basis set, the principal components (PCs). This transformation is achieved by the procedure described in Materials and Methods. If the variables are indeed correlated, then the first few PCs should capture most of the information in the original data. The variables in the case of X-ray data are the intensities at every diffraction angle and the interrelation can be seen from the correlation curves (Figure 4), where they increase or decrease with hydrolysis rates.

As can be seen in Figure 15b, there are points where the spectra intersect before changing the type of correlation with the hydrolysis rates. Division of the mean centered columns of the X^T matrix by the standard deviation results in a large variation in these parts (Figure 33 Appendix D) and would bias the estimation of the PCs (similar results were obtained for FC, data not shown). Therefore, no normalization with respect to the standard deviation was performed for computing the Z matrix. One of the reasons for standard deviation normalization is to make the variables dimensionless. However, since the X-ray data are of the same units and have already been normalized with respect to the area, standard deviation normalization is not required.

A plot of the singular values (Figure 19a) shows that there is one PC that accounts for 86% ($1.45^2/(1.45^2+0.41^2+….+0)$) of the variation in the case of Avicel and 96% for FC ($2.16^2/(2.16^2+0.30^2+…)$). As expected, the first principal component (Figure 19b) is basically a linear combination of the untreated cellulose and PASC spectra. The second PC seems to bring the main peak to the correct position in the reconstructed spectrum (not shown). The third and fourth PCs show similar importance. However, most of the variation in the data is captured by the first PC. The R^2 values of the reconstructed spectra from one PC and the original spectra are high (≥ 0.9 for Avicel and ≥ 0.96 for FC,

Table 14 in Appendix D), indicating a good fit. Also, the second PC scores do not have any correlation with the hydrolysis rates unlike the first PC scores which are linearly related to the hydrolysis rates (R^2 = 0.962 and 0.9553 for Avicel and FC respectively; 95% confidence intervals: slope – -0.1287 ± 0.0141 and -0.1507 ± 0.0196 for Avicel and FC respectively, intercept – 0.6830 ± 0.0836 and 0.8763 ± 0.1275 for Avicel and FC respectively; p-value for the full model was of the order of 10^{-10} for both Avicel and FC, thus the linear model is statistically significant) (Figure 20). This property will later be shown to be useful for hydrolysis rates prediction (section 4.3.5).

The correlation between the hydrolysis rates and the reconstructed spectra is stronger than that between the hydrolysis rates and the original spectra (Figure 17), where the correlation coefficients were ±0.98 and ±0.97 for Avicel and FC respectively. Using the reconstructed spectrum, the crystallinity index was calculated by equations (9) and (10). This is equivalent to regression of just the first PC scores. The crystallinity index values obtained were very close to those obtained with the whole data in section 4.3.2 (Figure 18 and Table 5, column 4); mean and maximum of the absolute difference between crystallinity indices from the two methods being: 0.67% and 1.68% for Avicel, 0.57% and 1.95% for FC.

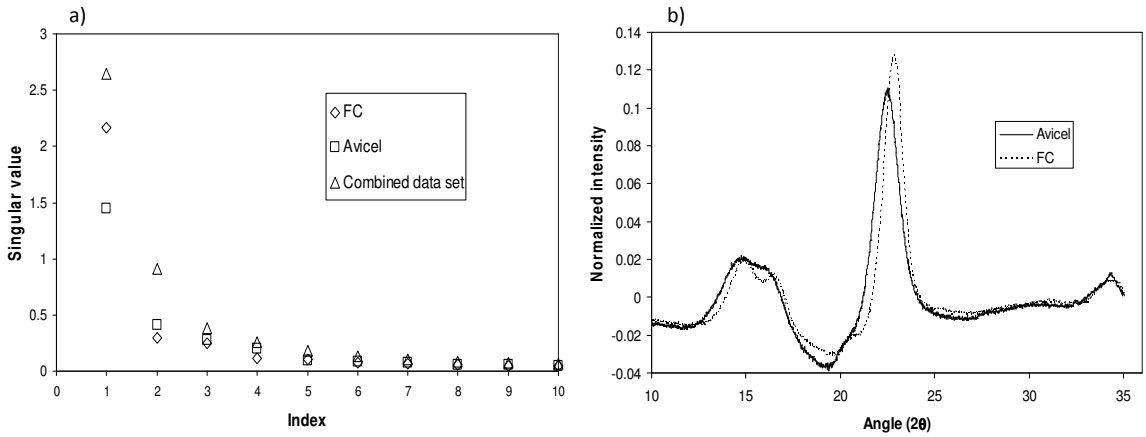
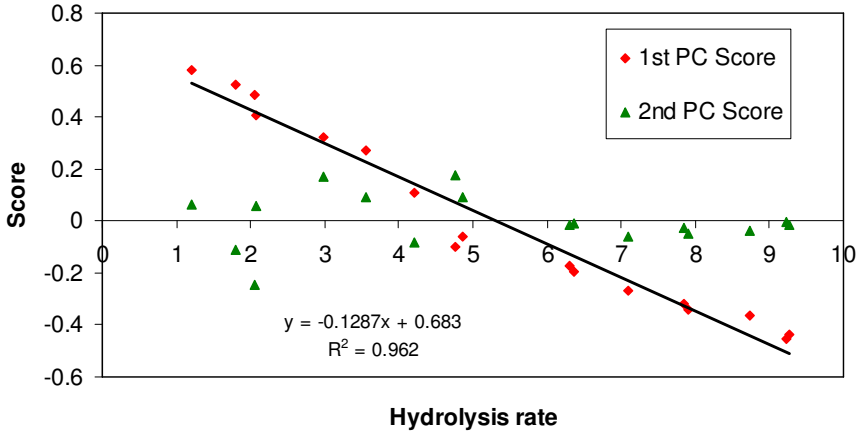


Figure 19. Plot of a) first ten singular values and b) first principal component of Avicel and FC data sets.

The R^2 values of the fit between the reconstructed spectra and the original spectra were also high (≥0.9 for Avicel and ≥0.96 for FC, Table 14 in Appendix D), showing that the spectra can be accurately reconstructed from just one PC. Since the spectra are reconstructed from only one PC, they are collinear in n dimensional space (n = 1500) and hence there will be no error term when using them in equation (9). PCA of the X-ray data shows that it is possible to collapse the entire spectra onto just one number (which is the score in the direction of the first PC) and subsequently use this to calculate the crystallinity index.

a)



b)

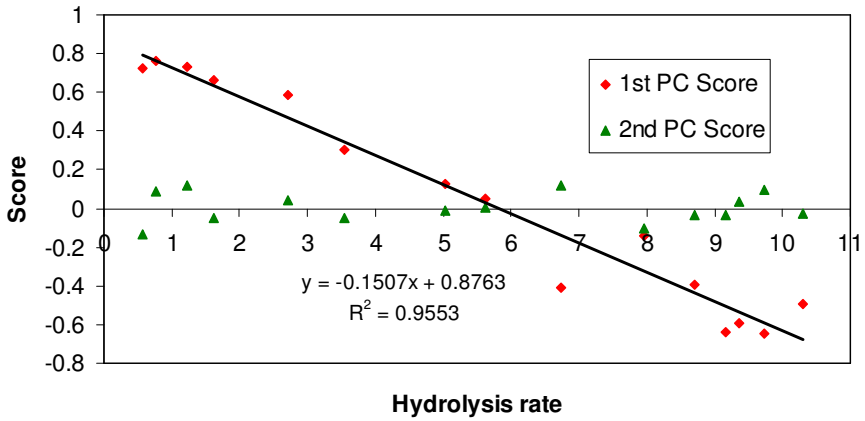


Figure 20. Plot of 1st PC and 2nd PC scores vs. hydrolysis rates for a) Avicel and b) FC. (Hydrolysis rates correspond to the amount of glucose produced in the first 2 min of the reaction with cellulases). The linear equations and the R^2 values of the fit between first PC scores and hydrolysis rates are also shown.

4.3.4 Validation of crystallinity calculation

The crystallinity calculation was validated by: 1) leave-one-out (LOO) validation method, and 2) using cellulose mixtures with varying fractions of untreated Avicel and amorphous cellulose. The idea behind LOO validation method is that the algorithm must be able to predict a particular sample's crystallinity index successfully when it is not included in the calculation of the PCs. The predicted crystallinity indices are then compared with those calculated when all the data are used in the PCA algorithm.
For LOO, each sample's spectrum was excluded once for the PCA. The spectrum was then reconstructed through its projection on the first PC and the corresponding crystallinity index was calculated by equations (9) and (10). The crystallinity values thus obtained were very close to the ones calculated with the whole data and PCA (Figure 18 and Table 5, column 6); mean and maximum of the absolute difference between crystallinity indices from the two methods (PCA and LOO) being: 0.21% and 1.32% for Avicel, 0.10% and 0.38% for FC.

To confirm the prediction ability of the new developed method, mixtures of untreated Avicel and amorphous Avicel (PASC) with various compositions were prepared (values refer to weight percentage): 80:20, 57:43, 40:60, 20:80. The calculated crystallinities (when the spectrum was not included in the computation of the PCs) were very similar to the theoretical ones (Table 6; Figure 34 in Appendix D), highlighting the power of the method; mean and maximum of the absolute difference between theoretical and calculated crystallinity indices were 1.64% and 2.99% respectively. This shows that the crystallinity index numbers are not artifacts but physically meaningful numbers representing the fraction of crystalline cellulose in a given sample of cellulose. In contrary, the peak height method did not predict well the crystallinity index (Table 6); mean and maximum of the absolute difference between theoretical and calculated crystallinity indices were 22.31% and 36.34% respectively.

Table 6. Theoretical and calculated crystallinity indices for various mixtures of untreated Avicel and amorphous cellulose, obtained by the method developed in this work, and peak height method resp.

Avicel fraction (%)	Theoretical Cri (%)[a]	Calculated Cri (%)[b]	R^2 (Fit between calculated and actual spectrum)	Theoretical Cri (%) (Segal)	Calculated Cri (%) (Segal)
80	49	49.75	0.981	77.72	86.81
57	34.2	33.98	0.970	52.62	71.45
40	27	29.99	0.961	48.55	73.51
20	12	14.59	0.953	18.46	54.80

[a] Theoretical Cri = (Avicel fraction*Cri_C + amorphous fraction*5), Cri_C = 60% (The cellulose sample used for preparing the samples was found to be not completely amorphous and had a calculated Cri of 5%).
[b] As the samples were prepared by mixing Avicel and amorphous cellulose, a mixture with a given crystallinity might not be as microscopically homogenous as a pure sample with the same crystallinity, giving rise to some errors in the crystallinity calculations.

4.3.5 Prediction of hydrolysis rates from X-ray data

The calculated crystallinities and the first PC scores are linear with respect to enzymatic hydrolysis rates (Figure 18 and Figure 20). Therefore, hydrolysis rates could theoretically be predicted from the calculated crystallinity indices or from the spectra themselves. Since crystallinity is just a property calculated from the spectra, we aimed at predicting the hydrolysis rates from the spectra itself. Doing so is not possible without any data transformation as the number of parameters (equal to the number of dimensions plus one constant) would far exceed the number of equations (equal to the number of samples). Through dimensionality reduction via PCA, we can overcome this problem. PCA of X-ray data shows that only one PC is sufficient to describe the X-ray data, and the hydrolysis rates are linear with respect to the first PC scores (Figure 20). It is interesting to note that the R^2 of the fit of hydrolysis rates with crystallinity (from PCA) is the same as that with the first PC scores (R^2 = 0.962 and 0.9553 for Avicel and FC respectively). Regression of reconstructed spectra in equations 9 and 10 is indeed equivalent to regressing just the first PC scores; since there is no error term involved due to only one

PC being used, the crystallinity indices are linear with respect to the first PC scores. A statistical test on the significance of the parameters shows that for principal component regression (PCR) (regression of hydrolysis rates vs. PC scores) also, one PC suffices (p-value <0.0001 for β_0 and β_1 of equation 18, and equal to 0.08 and 0.28 corresponding to β_2 for Avicel and FC respectively; p-values of 0.008 and 0.28 imply that β_2 is not statistically significant). PCR thus is able to accurately predict the enzymatic hydrolysis rates of cellulose samples using the X-ray spectra information.

The trend observed for the initial enzymatic hydrolysis rates of Avicel and FC may however be different for rates over longer reaction times, and should not be generalized to other cellulosic substrates without independent experiments. X-ray spectra of cellulose materials still have the power to estimate the degree of crystallinity of cellulose with consistency, and can give valuable information on the substrate and its susceptibility to enzymatic attack without actually performing any hydrolysis experiment.

A comment on the possible discrepancy between hydrated conditions for enzymatic reaction and dry cellulose samples for X-ray crystallography: enzymatic hydrolysis of cellulose necessarily occurs under hydrated conditions, whereas we employ dry (freeze-dried) samples for X-ray crystallography. The crystallinity index of hydrated samples can be measured as well, such as by the method of acid hydrolysis kinetics in boiling hydrochloric acid (Clarkin and Clesceri, 2002). Indeed, there are reports of differences in crystallinity index values of dry and hydrated cellulose from some sources as measured by ^{13}C NMR (Park et al., 2009). However, we do not expect the degree of hydration to have a major impact on our results. First, the method developed in this work has been validated with leave-one out cross validation and samples with known percentages crystalline and amorphous cellulose (section 4.3.4). Second, all the pretreated samples were subjected to the same freeze drying conditions, and thus fixing the reference crystallinity index value at 60% (Avicel) or 72% (FC) reduces systematic deviations. Last, it was found that freeze drying did not affect the Avicel crystallinity

index. Nevertheless, the question of the congruence of crystallinity indices measured by ^{13}C-NMR spectroscopy with dry or hydrated samples should be investigated further.

4.3.6 PCA of Avicel and FC spectra sets together

PCA on the combined spectra sets of Avicel and FC was done to investigate the possibility to describe the data sets with a limited number of PCs. The motivation for doing so was to develop a crystallinity index calculation tool which can hold for both types of substrate; given multiple cellulose I substrates and their spectra, it might be possible to describe all of them with a very few PCs. A first look at Figure 19b shows that the first PCs for the two substrates are similar. A plot of the singular values for the combined data set (Figure 19a) showed that although the first PC accounts for 86% of the variation ($2.65^2/(2.65^2+0.91^2+\ldots)$), two PCs might be required as $\sigma_2^2 = 0.83$, > mean($\sigma_1^2 + \sigma_2^2 + \ldots.\sigma_{32}^2$) = 0.26; σ_i denotes the ith singular value. The role of the second PC seemed to be to bring the peak of the 200 plane to the correct angle, e.g. from 22.7^o with one PC to 22.4^o as in the normalized spectrum for Avicel (Figure 21). The R^2 values of the fit between reconstructed spectra from one PC and the original spectra were much higher for FC ($0.93<R^2<0.99$; mean $R^2 = 0.97$) than Avicel ($0.85<R^2<0.99$; mean $R^2 = 0.93$). These R^2 values increased when the spectra were reconstructed from 2 PCs ($0.94<R^2<0.99$, mean $R^2 = 0.98$ for Avicel and $0.96<R^2<0.996$, mean $R^2 = 0.99$ for FC). For calculation of crystallinity index, the reconstructed spectra from two PCs were used in the following equations:

$$\mathrm{I_j}(2\theta)_r = f_{j1}\mathrm{I_P}(2\theta)_r + f_{j2}\mathrm{I_A}(2\theta)_r + f_{j3}\mathrm{I_{FC}}(2\theta)_r + \varepsilon \quad (19)$$

$$f_{j1} + f_{j2} + f_{j3} = 1 \quad (20)$$

The least square estimates of f_{j2} and f_{j3} - $\hat{f}_{j2}$ and $\hat{f}_{j3}$ were then used to calculate the crystallinity of the jth sample:

$$\mathrm{Cri_j} = \hat{f}_{j2} * \mathrm{Cri_A} + \hat{f}_{j3} * \mathrm{Cri_{FC}} \quad (21)$$

Cri_A (=60%) and Cri_{FC} (=72%) are the reference crystallinity indices of Avicel and FC respectively.

The crystallinity indices calculated with the combined data sets were very close to those with PCA on the individual data sets (Figure 35 in Appendix D). The mean and maximum of the absolute difference between the crystallinity indices from the two methods were 0.87%, 3.58% for Avicel and 3.66%, 6.84% for FC. The PCA on the combined data can thus successfully predict the degree of crystallinity of either Avicel or fibrous cellulose by projecting them onto two PCs: the first one captures the major variation in the spectrum while the second one primarily brings the 200 peak to the correct angle.

In order to predict hydrolysis rates from the combined data set, and have an expression applicable to both Avicel and FC, the hydrolysis rates were regressed to the principal component scores of the two PCs obtained from the combined data set. The high R^2 value (0.92) shows the good fit of the linear relation between the rates and the PC scores. The method outlined in this section can thus be used to predict the crystallinity and hydrolysis rates when two types of cellulose I substrates are at hand. Once the PCs have been determined, it is possible to calculate the degree of crystallinity and predict the hydrolysis rates of either type of cellulose I with any given crystallinity.

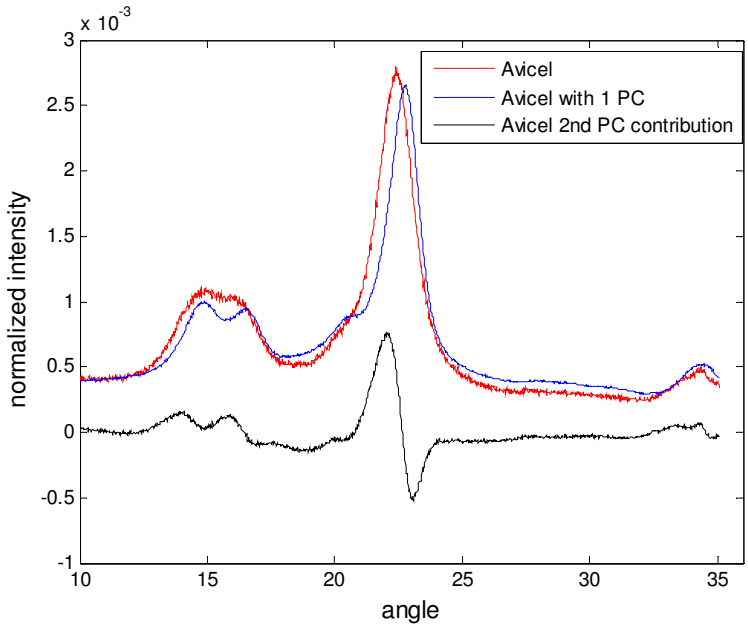


Figure 21. Normalized Avicel spectrum, reconstructed Avicel spectrum with one PC from the combined data set, and the contribution of the second PC to the Avicel spectrum.

4.4. Conclusions

A new method for calculating the crystallinity index of cellulose from X-ray diffraction data was developed and tested on samples of intermediate degrees of crystallinity. Dimensionality reduction of the normalized X-ray spectra revealed that they are highly concentrated along a single dimension and it is possible to collapse a spectrum onto one number. The crystallinity indices calculated by principal component analysis (PCA) with one principal component were similar to those obtained by regressing the whole spectrum. The method was validated by the leave-one-out method, and the calculated crystallinities of cellulose mixtures prepared with varying ratios of Avicel and amorphous cellulose were shown to be consistent with the theoretical values. The data set produced in this work can now be used to calculate the crystallinity index of any Avicel and fibrous cellulose sample, which may have relevance for evaluating the efficacy of a pretreatment

method, understanding crystallinity changes over the course of enzymatic hydrolysis, understanding the relationship of crystallinity to other properties such as degree of polymerization (DP), or any other application where crystallinity plays an important role. Recently, it was successfully applied to Cel7A CBD pretreated cellulose (Hall et al., 2011), where CBD pretreated samples were shown to be having reduced crystallinity. It was also applied to partially converted Avicel, and was in agreement with solid state 13C NMR in showing constancy of cellulose crystallinity with conversion (Chapter 3). PCA also makes possible the accurate prediction of hydrolysis rates from X-ray spectra data sets, as calculated crystallinity indexes were found to be linearly related with the corresponding hydrolysis rates.

The crystallinity index calculation method presented could additionally be tested with cellulosic substrates other than Avicel and fibrous cellulose. Though the results might give different trends, the overall framework will still be applicable – data normalization, calculation of crystallinity index from the normalized data, dimensionality reduction through PCA, calculation of crystallinity by PCA, and regression of hydrolysis rates if possible.

CHAPTER 5

COMPUTATIONAL ANALYSIS OF THE PROTEIN SEQUENCE SPACE: A NEW METHOD TO IDENTIFY TARGET MUTATIONS

(Experimental work on Old yellow enzymes was done by Dr. Yanto Yanto and Jonathan Park)

The finding that the cellulose binding domain (CBD) of Cel7A can reduce cellulose crystallinity upon pretreatment (Hall et al., 2011), and also impart thermostability to the intact Cel7A (Hall et al., in press), shows that it is a promising target for protein engineering. As far as protein engineering is concerned, the assay required to check for crystallinity reduction and a further increase in hydrolysis rates requires incubation times of 15 hours. Thus, any approach akin to directed evolution requiring a high throughput assay is infeasible. Rational design driven by structure-function relationship may be applied here to engineer the CBD residues interacting with the cellulose surface. However, knowledge is limited as Cel7A CBD mutations have been investigated to study adsorption effects only (Carrard and Linder, 1999; Linder et al., 1995a; Linder et al., 1995b). Also, since only one three dimensional structure is available for the family I CBDs (http://www.cazy.org/CBM1_structure.html, as of August 10, 2011), structural comparison with other members of the family is not possible. The third wave of protein engineering involving statistical tools to relate sequences and functionality requires an initial data set which also entails significant experimental effort. The generation of the list of suggested mutations thus has to be guided by a systematic method. SCHEMA (Meyer et al., 2003), Rosetta (Siegel et al., 2010), and CASTing (Reetz and Carballeira, 2007) have shown promise, but all of these require structural knowledge.

A method utilizing patterns in a protein family's sequence has been developed using principal component analysis (PCA). The underlying idea is to identify patterns in a protein family's sequences, and then look for changes (mutations) in our protein of interest such that the new variant is closer to the underlying low-dimensional manifold of the sequences. This low-dimensional topology is identified by subjecting the protein family sequences to principal component analysis (PCA). Non-negative matrix factorization (NMF) (Lee and Seung, 1999), another linear dimensionality reduction technique that explores the linear sub-space of a high dimensional data set was also tested. ISOMAP (Iso-lateral mapping), a non-linear dimensionality reduction technique (Tenenbaum et al., 2000), was found to explain the data in a lower number of dimensions, but since the mapping back to sequence space is not easy, its applicability is limited.

To gauge the performance of the PCA based method, the family of sequences of proteinase K was run through the PCA algorithm, and the suggested mutations at the 24 positions were compared with those in Liao et al. (2007). The family I of CBDs was also analyzed and the suggested mutations will be used by Yuzhi Kang in the Bommarius lab on Cel7A CBD. Mutations suggested for Old Yellow enzymes have also been tried (in collaboration with Dr. Yanto Yanto and Jonathan Park), but no improvement in activity was observed.

5.1 Methodology

Figure 22 shows the approach for identifying the mutations

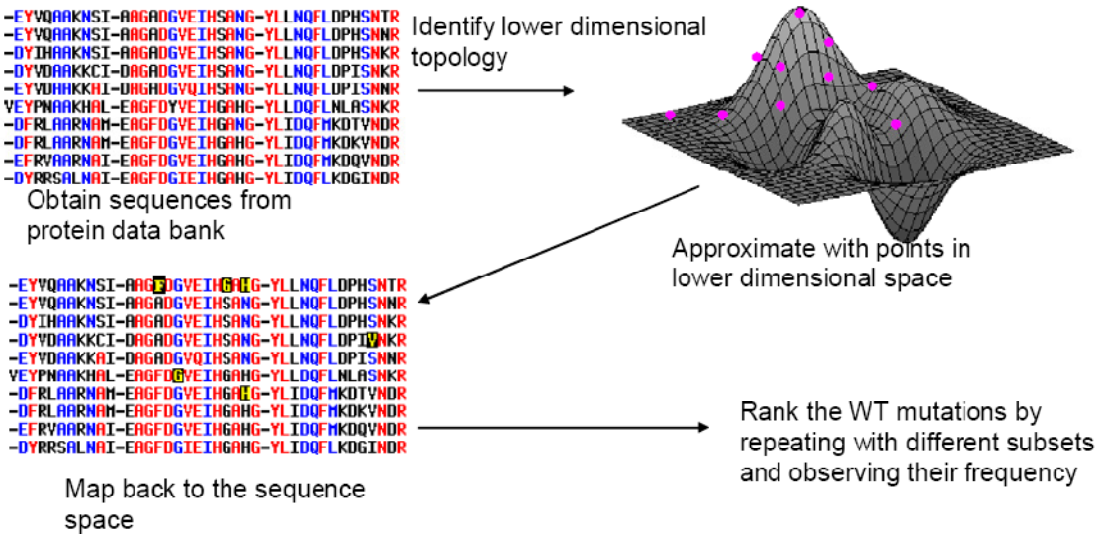


Figure 22. Illustration of the working of PCA based sequence analysis to identify target mutations. In yellow are shown the mutations upon mapping back to sequence space (note: this is just an illustration, not a real example).

Step 1: Collect protein sequences from the protein data bank (PDB), align them using an alignment tool like Clustal W (Thompson et al., 1994). Residue at a position is defined based on the sequence alignment.

Step 2: Subject the sequence data set to PCA, and reconstruct the protein sequence of interest with a limited number of principal components. Positions that upon reconstruction have a residue different than the one present originally are the positions of interest for mutating, and the suggested residues from the PCA are the target residues for substitution.

Step 3: Repeat the PCA and sequence mapping (reconstruction) with different data sets of the collected sequences, and rank the mutations based on the weight of the suggested residue and its frequency of occurrence. Gaps are excluded when considering mutations.

There are two parameters that the user has to define. These are explained below:
- Window sizes: As stated above, different sub-sets of the sequences are used to identify mutations. The reason for this is to avoid biasing of the suggested residue by the closest homologues. So, our protein of interest (say, protein number 1), is selected with (say) *n* other proteins from the sequences. This sub set then cascades down by one in the table, and so on. The number *n* is the window size. This procedure can be repeated with multiple window sizes, the range of which is up to the user.

To test for variance (standard deviation) in the weighted frequencies of mutations, any two sequences can be swapped or excluded to repeat the PCA. Frequency is the number of times a mutation is observed for a chosen subset. Weighting is explained in section 5.1.3 (Reconstruction of protein sequences).

- Percentage cut-off for excluding positions having gaps: Upon alignment, there will be some positions having gaps in many of the sequences. These positions are considered as having no residue, or in mathematical terms, having lack of information for the data set. To avoid biasing the results from these positions with lack of information, only those positions with a certain percentage of information are included in the PCA. The cut-off is to be set by the user.

The number of principal components can also be set as a parameter, but it is suggested that these be varied from 1 to the maximum number (in this work, it will be the number of data points/sequences).

5.1.2 Feature space

To represent the proteins mathematically, each position was assigned twenty dimensions, corresponding to each amino acid residue. The order of assignment among the twenty amino acid residues does not affect the computation. If an amino acid is present at a particular position in a sequence, then the dimension corresponding to that amino acid residue at that position is assigned a value of 1, if not then 0. For example, the amino acid block AST will look like:

[1 0 0…..0| 0 0…..1 0 0 0|0 0 0…..1 0 0 0]

when dimensions are assigned alphabetically – A, C, D,… Y for alanine, cystine, aspartic acid, … tyrosine.

Every position could have alternatively been assigned a number based on a physiological property, in which case the results would have been determined or biased rather by that property. The selection of the feature space chosen in this work ensures that the principal components and the reconstructed sequences are driven solely by occurrence of amino acid residues. It also has the advantage of being linearly separable if at some point classifying tools such support vector machines are to be applied (Dubey et al., 2005).

The data matrix will be a sparse. The dimensions with all zeros or all ones will have no error upon reconstruction (for proof see Appendix E). In the computation of this work, all dimensions with zeros were removed to avoid excessive memory usage in MATLAB® (The Mathworks Inc. R2008b).

5.1.3 Reconstruction of protein sequences

The protein sequence data, expressed as given below, was subject to principal component analysis. More details of PCA can be found in Jolliffe (2002).

$$X^T = \begin{pmatrix} x_{11} & x_{12} & \ldots\ldots\ldots x_{1p} \\ . & & \\ . & & \\ x_{n1} & x_{n2} & \ldots\ldots\ldots x_{np} \end{pmatrix}_{n \times p} \quad (1)$$

where n is the number of sequences, p is the number of dimensions = 20*number of positions in alignment.

From this matrix, the dimensions with all zeros, or all ones were eliminated for computation of the principal components. To find the principal components (vectors in the n dimensional space along which the data might be concentrated), singular value decomposition (SVD) of the Z matrix (given below) is performed. The Z matrix is obtained by mean centering the columns of X^T. Standard deviation normalization is not performed because of the presence of completely conserved positions (a zero divided by zero situation will be encountered in that case). This is also the case with X-ray data as mentioned in Chapter 4.

SVD of the Z matrix will give the principal components through -:

$$Z = U_{p \times p} S_{p \times n} V^T_{n \times n} \quad (2)$$

U and V are matrices consisting of orthonormal vectors.

$$S = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 \ldots\ldots & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 \\ . & & . & & \\ . & & & . & \\ 0 & & & & \sigma_n \\ 0 & & & & 0 \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ 0 & & & & 0 \end{pmatrix}_{p \times n} \quad (3)$$

σ's are the singular values.

The Z and X matrix can be reconstructed (Z_r and X_r) from the reduced dimensional score matrix Z_{tf}:

$$Z_{tf} = U^T_L Z = S_L V^T_L \quad (4)$$

$$Z_r = U_L Z_{tf} = U_L U^T_L Z \quad (5)$$

$$(X_r)_{ij} = \bar{x}_i + (Z_r)_{ij} \quad (6)$$

where $\bar{x}_i$ is the mean of the i^{th} column in X^T.

For any position in the original data matrix we have a 20 dimensional array – [0 0 0 0 …. 1 0 0 0] where there is a 1 for the residue present at that position and 0 otherwise. In the reconstructed matrix, this may not be the case: [0 0 0 0 …. 0.3 0.45 0.25 0]. The value highest in magnitude (closest to 1) is then rounded off to 1 and others to 0 to approximate the residue present in the reconstructed array. The value in the reconstructed matrix is the weighting; in the above example the weighting is 0.45. When the mutations are scored with different subsets of the sequences, this weighting is also taken into consideration.

Upon reconstruction the twenty elements corresponding to a position always sum to unity, except in some cases where there is no residue present for some sequences (gaps). Residues suggested at gaps are not considered while counting mutations.

5.1.4 Properties of the method

Some of the properties of the method that will become clearer through its applications to certain data sets in the next few sub-sections are -:

1. The method tells which positions to mutate and what residues to mutate them to, in order of their ranking.
2. PCA involves eigenvalue decomposition of the covariance matrix. Therefore, covariations and coevolution of residues is accounted for. This is more than just pair-wise correlation analysis.

3. The target residue suggested at a position is within the library. In other words, it will not suggest a residue that is outside those (at any given position) in the family of sequences. This also implies that there will be no mutation suggested at completely conserved positions; a mathematical proof of this can be found in Appendix E.
4. It requires no activity/functionality data as it is uses only the protein sequences. In data-mining or machine learning terminology, it is a case of unsupervised learning. As stated in conclusions and perspectives, activity can be incorporated into the method.

All computations were performed in MATLAB® (The Mathworks Inc. R2008b).

5.2 Test case – Proteinase K

5.2.1 Application to proteinase K data

Liao et al. (2007) selected twenty four mutations (twenty four target residues at twenty four positions) in proteinase K based on literature reports. Using various linear regression models, they identified ten positive mutations and were able to achieve a twenty fold improvement in the activity after three rounds. Their procedure is shown in Figure 23.

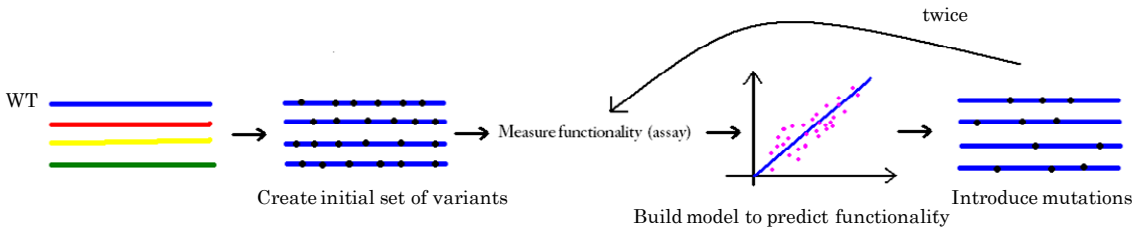


Figure 23. Framework of Liao et al. (2007)

The performance of the PCA method was checked by applying the PCA method to the 57 sequences used by Liao et al. (2007). The top ranked suggested mutations were then compared with those in the original work. Parameter settings and data size are shown in Table 7.

Table 7. Data set size and parameter settings for proteinase-k data.

Data size*	57
Percentage cut-off	70%
Window size	25

*Sequences had greater than 30% identity with respect to the proteinase K of interest.

The top twelve ranked mutations are shown in Table 8, and their scores (weighted frequencies) with standard deviations in Figure 24. Of the ten positive mutations in Liao et al. (2007), four appear in the top twelve ranked mutations, and three in the top six. Some mutations have an unknown effect because the residues suggested by the PCA method at those positions are different from the ones chosen in the original work. The number of positive mutations in Table 8 might therefore be more than four.

Table 8. Ranking of mutations (first 12) and their effects.

Rank	Mutation	Effect
1	G293A	Strong positive
2	S273T	Positive
3	V167I	Negative
4	N95G	?
5	S123T	?
6	I132V	Positive
7	K208S	?
8	S337A	?
9	K332G	?
10	M145Q	?
11	S107Q	?
12	L180I	Positive

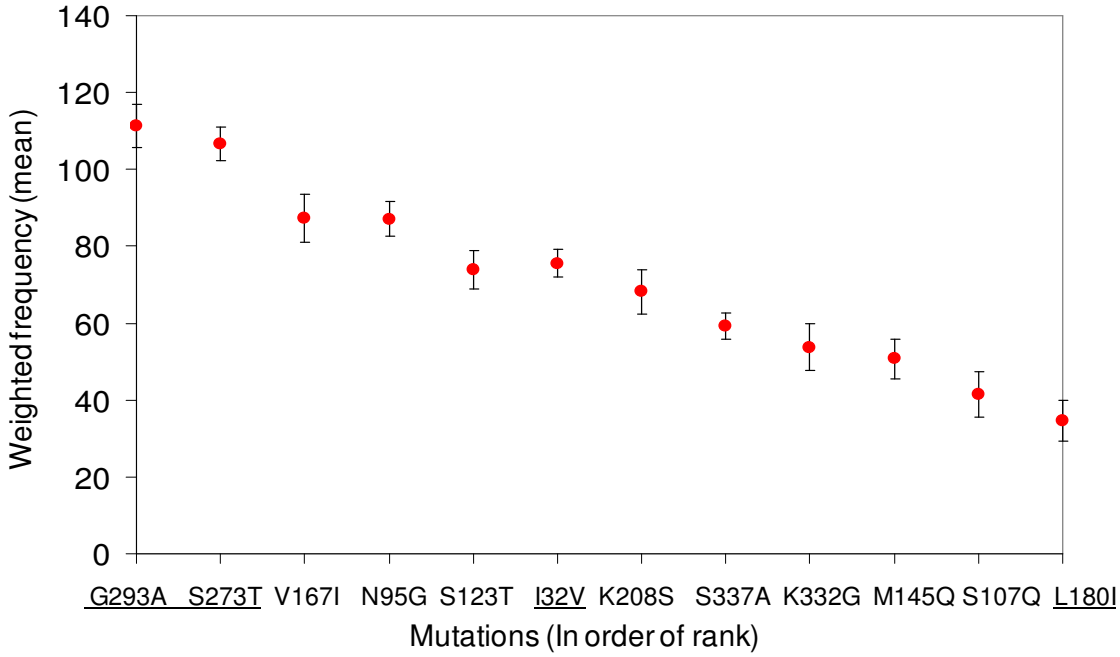


Figure 24. Weighted frequencies of the first 12 mutations. Positive mutations are underlined.

5.2.2 Comparison with consensus approach

According to the consensus method of chosing a mutation, the residue at the position of interest is mutated to the one that is present in the majority in the family of the sequences (Steipe et al., 1994). The comparison is shown in Table 9. Outside the first three mutations, there is no pattern followed as far as the consensus residue is concerned. The relation between a consensus residue and a PCA suggested mutation can be understood from the equation used for reconstruction (mapping back to the protein sequence) (equation 6).

$$(X_r)_{ij} = \bar{x}_i + (Z_r)_{ij} \qquad (6)$$

First term on the right hand side is the mean, which is representative of the frequency of occurrence of a certain residue at that position. If there is a very commonly occuring

residue, then it will dominate the reconstruction, and therefore the suggested residue. However, if there is no majority in the residues at a position, then the second term on the right-hand side, the correlation term, will determine the suggested residue. The PCA suggested residue constitutes therefore a balance between the two.

Table 9. Comparison of PCA mutations with those suggested by the consensus approach (most commonly occurring, in parentheses is the fraction of sequences having that residue), and the BLOSUM 62 matrix residues (scores in parentheses).

Rank	Mutation	Effect	Most common?	BLOSUM 62 score	BLOSUM 62 residue
1	G293A	Strong positive	Yes (0.88)	0	A, S, N (0)
2	S273T	Positive	Yes (0.7)	1	N, A, T (1)
3	V167I	Negative	Yes (0.5)	3	I (3)
4	N95G	?	No (0.43)	0	L (2)
5	S123T	?	Yes (0.47)	1	N, A, T (1)
6	I132V	Positive	No (0.4)	3	V (3)
7	K208S	?	No (0.21)	0	R (2)
8	S337A	?	Yes (0.35)	1	N, A, T (1)
9	K332G	?	Yes (0.47)	-2	R (2)
10	M145Q	?	No (0.14)	0	L (2)
11	S107Q	?	Yes (0.47)	0	N, A, T (1)
12	L180I	Positive	No (0.33)	2	I, M (2)

The mutations were also compared with the BLOSUM 62 matrix, which is a substitution matrix used to score alignment between proteins (Henikoff and Henikoff, 1992). Out of the twelve mutations shown, only two clearly matched with the BLOSUM 62 suggested residue (V167I and I132V), and four other mutations were from one of the BLOSUM 62 suggested residues (G293A, S273T, S123T and L180I) (Table 9). It is not surprising that some of the PCA mutations match the residues from this matrix because the input to the PCA method is the sequence alignment which itself uses the BLOSUM matrices.

5.3 Test case – Old Yellow Enzymes

Old Yellow Enzymes are flavoproteins catalyzing the reduction of activated alkenes and can produce up to two chiral centers in a stereospecific manner (Hall et al., 2010b). One hundred and thirty two mutations were identified in the ene-reductase from *Yersinia bercovieri* (Yers-ER) *via* PCA (Figure 25). The data set consisted of twenty eight sequences, with 22 to 76% identity with respect to Yers-ER. The percentage cut-off was fixed at 70%, whereas the window size was varied from 16 to 24.

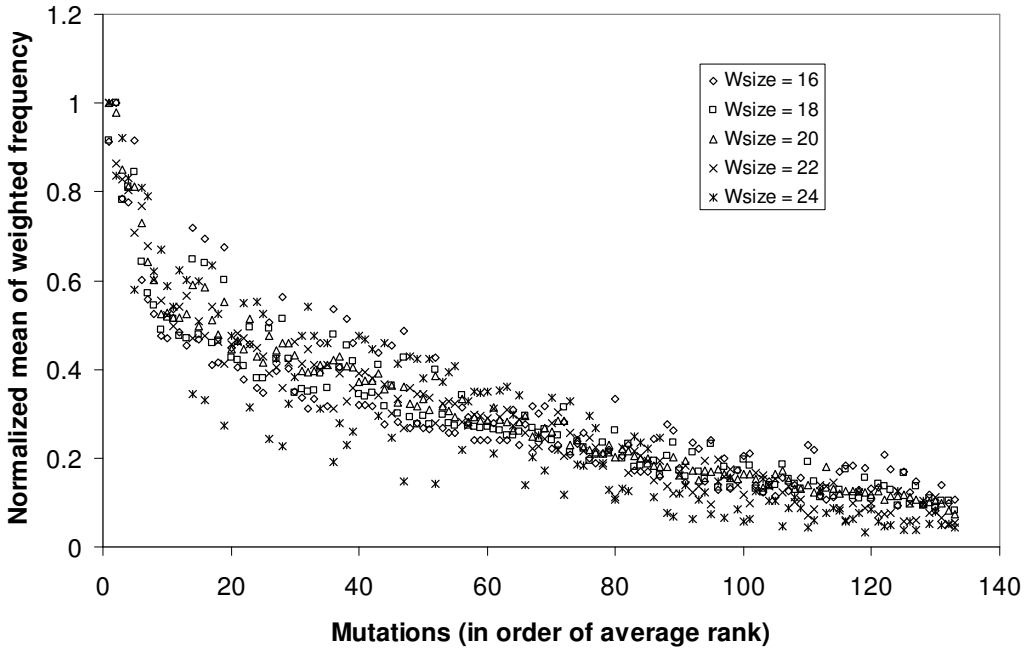


Figure 25. Scores (normalized mean of weighted frequency) of mutations for various window sizes (Wsize).

Of the one hundred and thirty two mutations suggested by PCA, the ones present in the first and second shell of the bound flavin molecule were chosen. The first- and second-shell residue is defined as any residue having any part within 4 or 8 Å of any part of the

flavin molecule. The structural analysis was performed in PyMOL (The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC.). Of the twelve mutations identified, four showed improvement over the wild-type, four showed comparable activity, two were undetermined, and only two showed a drastic decrease (Table 10). No change in enantioselectivity was observed.

Table 10. Activities of twelve variants containing mutations in the first and second shell of flavin molecule. Color coding: black – comparable to WT, blue – greater than WT, red – less than WT.

		Specific Activity [U/mg]			
Avg Rank	**Mutant**	**2-cyclohexen1-one**		**Ketoisophorone**	
-	WT	2.76	-	5.95	-
25.2	G348P	2.53	91.70%	5.87	98.60%
26	A72T	0.49	17.80%	1.28	21.50%
33.4	I235L	3.7	134.10%	7.14	120.00%
35.2	H185N	0.05	1.80%	0.18	3.00%
38.4	I271M	2.72	98.60%	4.8	80.70%
45.4	L181I	1.98	71.70%	4.41	74.10%
59.8	T350S	4.04	146.40%	7.7	129.40%
62	T57S	NA	NA	NA	NA
67	A183S	4.41	159.80%	8.42	141.50%
83	A303R	2.82	102.20%	6.73	113.10%
140.2	T305D	NA	NA	NA	NA
144.4	P346Y	2.61	94.60%	5.38	90.40%

NA – not available

To investigate the effect of mutations near the substrate, of the one hundred and thirty two mutations, the ones present in the first and second shell of the bound 2-cyclohexen1-one in Yers-ER were tested for their influence on activities of Yers-ER on 2-cyclohexen1-one and ketoisophorone. No improvement over the WT was observed (Figure 26) (three variants - T133F, T133S, and V134L were not asssayed with

consistency and most likely give activity less than the WT). When the amount of flavin bound is considered, V105T and T132V show activities comparable to the WT for cyclohexenone (Figure 26).

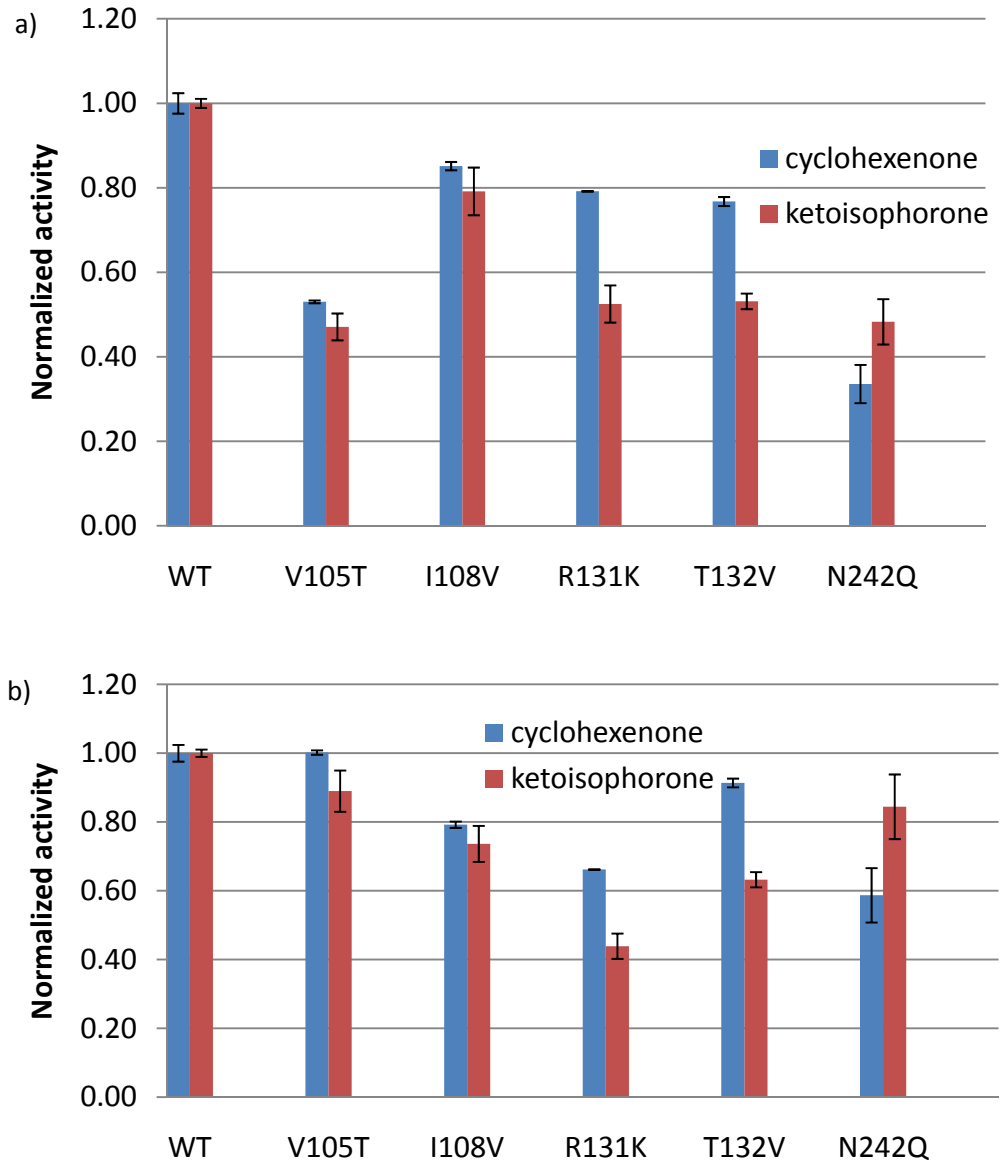


Figure 26. a) Activities of variants containing mutations in the first and second shell of the bound cyclohexenone, b) activities based on flavin occupancy.

5.4 Application to family I of cellulose binding domains

5.4.1 Identification of mutations

A BLAST search was carried out on the Cel7A CBD sequence. 100 hits were obtained out of which 49 were collected, based on screening repetitions, and selecting only those that were reported in publications. The sequences and organism sources are given in Appendix F. Eight of these were putative sequences, which were excluded from the analysis. With a window size of 25 and percentage cut-off of 70%, the ranking of mutations is shown in Table 11 (one run only, no variance added by sequence exchange or exclusion). An interesting observation is that S3T has the highest frequency but due to its lower weighting compared to S14T, it is not ranked first. Similary, Q26K occurs more often than G22P, Y5W and Y13W but with a lower weighting.

Table 11. Ranking of mutations in Cel7A CBD (ordered according to the weighted frequency).

Mutation	Frequency	Average weighting	Weighted frequency
S14T	34	0.61	20.85
S3T	43	0.33	14.22
V18T	35	0.36	12.70
G22P	17	0.58	9.82
Y5W	16	0.54	8.68
Y13W	13	0.63	8.18
Q26K	20	0.31	6.22
V18T	9	0.35	3.18
T23Y	8	0.34	2.71
T23A	8	0.30	2.40
H4L	4	0.35	1.40
S3A	5	0.27	1.33
I11S	3	0.40	1.21
Q26T	3	0.24	0.71

Standard deviations of the weighted frequencies in the first seven mutations were investigated (first seven were chosen because of their high frequency compared to the remaining ones), and results are shown in Figure 27. Mutations ranked 4 and 5 (G22P, and Y5W) show comparable means, and mutation 6 (Y13W) has a mean higher than them but with a high standard deviation.

These mutations are now candidates for studies related to reduction in cellulose crystallinity upon pretreatment with Cel7A CBD (and a simultaneous increase in rates), or thermostability improvement.

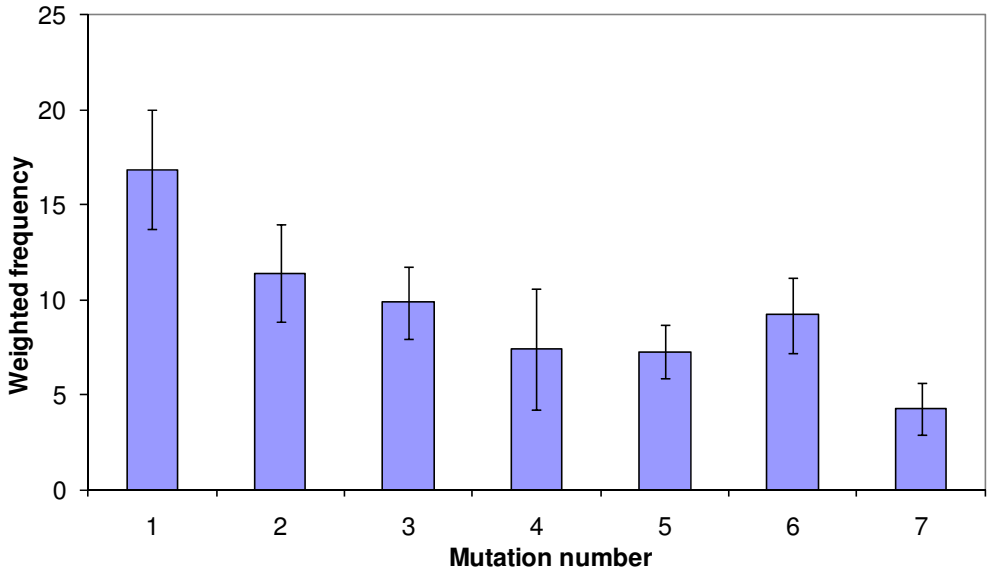


Figure 27. Weighted frequencies and their standard deviations in the different mutations, numbered according to Table 11**.**

5.4.2 Analysis of covariation in the library

To get an idea of how the interactions/correlations between residues at different positions might be important, two tests were done: 1) scrambling the residues position-wise and then checking for mutations suggested in the Cel7A CBD, 2) removing the information in

one half (positions 1 – 18 or 19 – 36) by making the entries zeros in the corresponding dimensions to check for mutations suggested in the other half.

As scrambling is random, every time we scramble, we are likely to get a different result. For one run, the results are shown in Table 12. Of the top nine mutations only five are captured, that too with a jumbled order, and low frequency and weighting.

When information in the first half of the sequence is removed, only Q26K shows up as the suggested mutation; implying that the mutations at positions 22, and 23 are lost. Ranking after removal of information in the second half is able to capture most of the mutations, although in a different order, and ranking (Table 13). As explained before in section 5.2.2 for proteinase K, the residue at a given position reconstructed from PCA has two main determining factors – the average residue, and the correlations with other positions (equation 6). It is clear that there is a strong correlation term as seen with the difference in results between Table 11 and Table 12 & Table 13.

.

Table 12. Mutations in Cel7A CBD for scrambled sequences.

Mutation	Frequency	Average weighting	Weighted frequency
V18T	13	0.30	3.90
S3T	12	0.29	3.48
S3A	8	0.32	2.56
G22P	3	0.54	1.61
H4K	4	0.28	1.12
S14T	1	0.51	0.51
H4V	2	0.25	0.50
V27T	1	0.33	0.33

Table 13. Mutations at positions 1 – 18 in Cel7A CBD when information at positions 19 – 36 is removed.

Mutation	Frequency	Average weighting	Weighted frequency
S3T	43	0.53	22.67
Y13W	10	0.64	6.44
S14T	8	0.58	4.63
Y5W	8	0.56	4.45
V18T	10	0.32	3.17
V18A	10	0.26	2.62
H4L	9	0.25	2.24
H4V	5	0.40	2.02
V18N	2	0.30	0.60

5.5 Non-negative matrix factorization

Non-negative matrix factorization (NMF) identifies the lower-dimensional linear sub-space of a data set with the constraints that the feature vectors and their scores are non-negative (Lee and Seung, 1999). Singular value decomposition (SVD) for PCA contains negative entries and thus is difficult to interpret for non-negative data sets such as images, texts, and protein sequences. NMF is able to achieve a parts based representation through the different feature vectors, e.g., a document is made up of various terms with

different frequencies. NMF is particularly useful in the case where the aim is clustering. The feature vectors (dimensions) in this case correspond to the clusters.

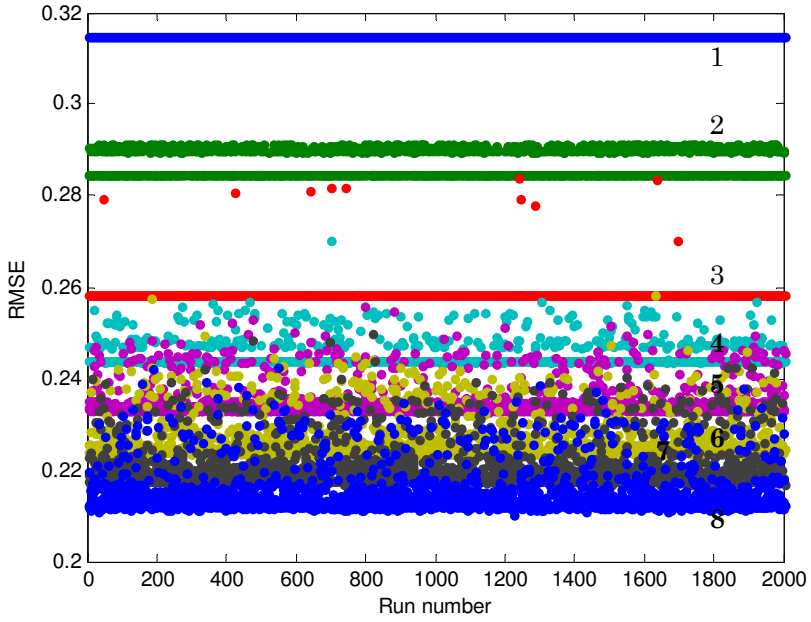
$$\min_{\mathbf{W},\mathbf{H}} f(\mathbf{W},\mathbf{H}) \equiv \frac{1}{2}\left\|\mathbf{A}-\mathbf{WH}\right\|_F^2 \qquad s.t.\ \mathbf{W},\mathbf{H} \geq 0 \tag{7}$$

A is the data matrix, W is the matrix consisting of the feature vectors, H is the loading/score matrix. For a positive definite D,

$$\mathbf{WH} = \mathbf{WDD}^{-1}\mathbf{H} \tag{8}$$

One of the shortcomings of NMF is the non-uniqueness of the feature vectors as seen in equation 8. In cases where uniqueness of the feature vectors is not a concern, e.g., clustering, classification, image cleaning, etc., NMF is useful. The most critical step in identification of mutations in the methodology developed in this work is the reconstruction or mapping back of the low-dimensional data to the sequence space. Figure 28 shows the root mean square error (RMSE) for CBD famliy I and OYE data sets. It is clear from the RMSE results that the degenerate solutions are obtained. The reconstruction was found to be degenerate too.

a)



b)

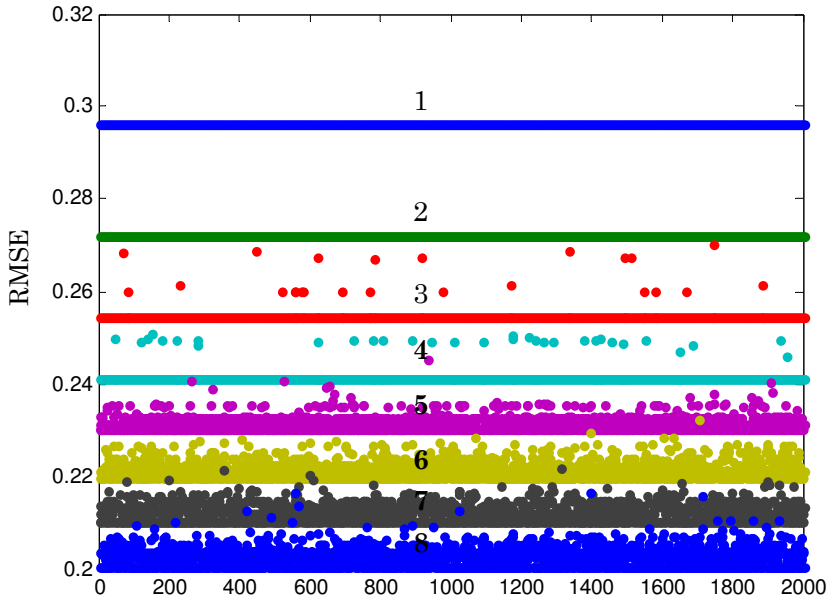


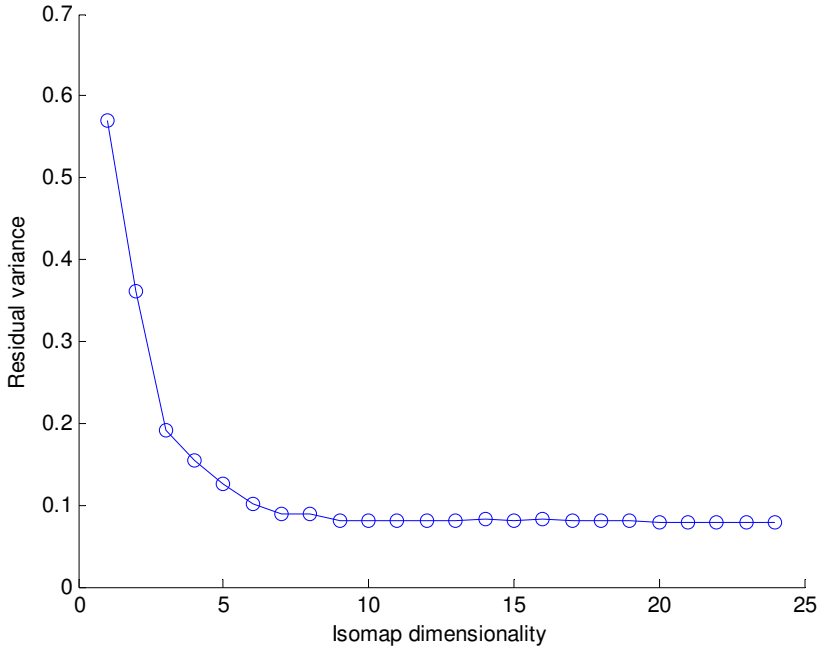
Figure 28. NMF RMSE for different runs - a) family I of CBDs, b) OYE data set. The number of dimensions are shown next to the RMSE value.

The current state of the art for NMF is not applicable for identification of target mutations in the protein sequence space. Perhaps, a different feature space that ensures uniqueness of NMF solutions, or a methodology non restricted by the uniqueness of matrix decomposition will be able to utilize NMF in future.

5.6 Non-linear dimensionality reduction

To explore the applicability of non-linear dimensionality reduction methods for identifying target mutations, Isomap was tested. Isomap (Tenenbaum et al., 2000), determines the low dimensional embedding by preserving the geometric distances of the original space. The number of dimensions need to be specified by the user, and so does the number of nearest neighbors or the fixed radius ε. Isomap is able to capture the variance in a smaller number of dimensions (Figure 29).

a)



b)

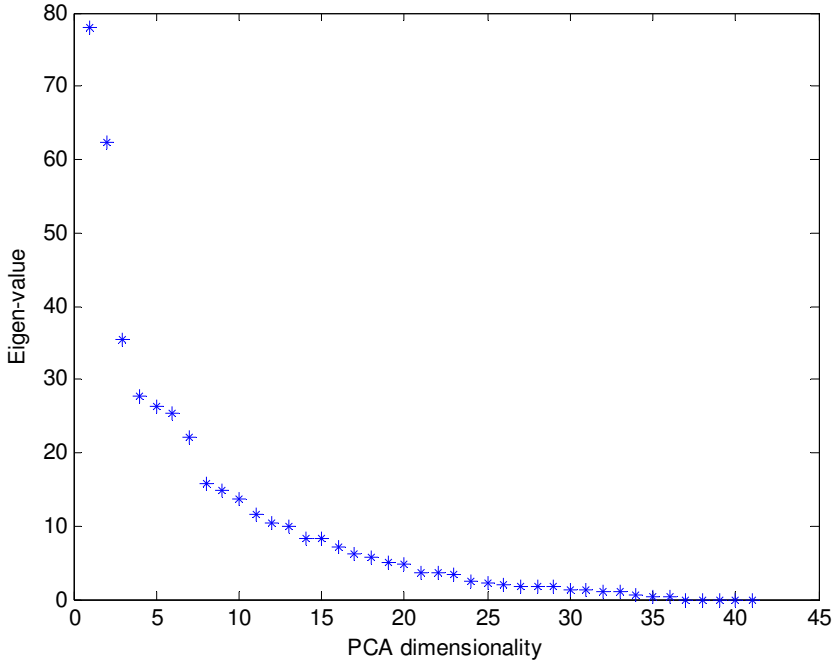


Figure 29. a) Residual variance for Isomap, and b) Scree plot for PCA, when applied to family I CBD data set.

Unlike PCA or NMF, Isomap and other non-linear dimensionality reduction methods do not output a set of basis vectors due to their non-linear nature. Reconstruction (called the pre-image problem in machine learning) then is not trivial. The current pre-image methods assume that any reconstructed point is within the span of its neighbors and requires singular value decomposition (Kwok and Tsang, 2004), thus involving something that is not any different from PCA. Non-linear methods such as LLE (local linear embedding) have been applied for protein classification (Wang et al., 2005), but the pre-image problem makes these methods difficult to apply in the framework developed in this work. In the case of available activity or *y* data, these methods can be very useful for classsification or regression.

5.7 Conclusions

A new method based on principal component analysis was developed to utilize the underlying pattern in the sequences of a protein family, and suggest mutations. The differences between residues at specific positions in the original sequences, and the ones closer to the landscape, are the target mutations. In the absence of a high throughput assay, the only surrogate for a library of active protein sequences are the homologs obtained from the protein data bank. Since activity data is not utilized in the developed PCA method, there might not exist a strong correlation between the ranking and the performance. However, as seen with proteinase K, the method is able to pick up beneficial mutations with a reasonable degree of success. The method can also be useful if the aim is to identify positions tolerant to mutations. In the case of Old Yellow Enzymes, even though there was no significant improvement over the WT activity, a major fraction of the variants were shown to have comparable activities. This has to be confirmed with a negative control study where random mutations are made at the chosen positions.

Non-negative matrix factorization (NMF) and Isomap were also tested on protein sequence data but were found to have limitations. NMF suffers from degeneracy whereas Isomap has limited applicability due to the pre-image problem. When suitable amounts of activity data become available for regression or classification, these methods should be investigated further as unique feature vectors or reconstruction (pre-image) may not be required. In this light, PCA, NMF, Isomap or other dimensionality reduction methods can be applied to any large library of assayed variants generated by directed evolution or rational design.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

The precipitous decline in the rates of the enzymatic hydrolysis of cellulose is one of the major limitations to the commercialization of second generation biofuel. Understanding the causes behind the decelerating rates has been challenging due to the interplay of many enzyme- and substrate-related factors. In this thesis, published kinetic modeling works on cellulose biohydrolysis were critically reviewed, and incisive kinetic studies were carried out to identify and quantify the various rate limitations. Cellulose crystallinity, a major rate governing property, which suffers from measurement and calculation inconsistencies, was quantified accurately and consistently using a method using multivariate statistical analysis.

Engineering cellulases for higher activity and thermostability, and cellulose binding domains for cellulose crystallinity disruption and thermostability is an attractive route to enhance cellulose hydrolysis rates. However, in the absence of high throughput assays, and lack of knowledge on the role of specific amino acid residues, mutations have to be picked judiciously. To tackle this issue, a method based on principal component analysis was developed and tested on literature as well experimental data.

6.1 Identifying rate limitations in the enzymatic hydrolysis of cellulose

The published kinetic models and experimental studies in the literature point to various substrate- and enzyme- related properties affecting cellulose hydrolysis rates along conversion – accessibility, intrinsic reactivity, enzyme inactivation, product inhibition, jamming, clogging, increase in cellulose crystallinity, decrease in enzyme synergism, fractal nature of the substrate, and depletion of chain ends for cellobiohydrolases. The plethora of factors and diversity in the kinds of substrates makes it challenging to

accurately identify and quantify the dominant factors. There exists no single study that has unequivocally pinpointed the phenomena behind the decreasing rates. Models based on Michaelis-Menten kinetics and those having empirical factors are expected to mask the real underlying factors responsible for rate retardation.

In this thesis, it was shown that the rate limiting phenomena that were either not occurring or could be avoided are - changes in average cellulose chain length (previous degree of polymerization experiments by Dr. Mélanie Hall with cellulase mixture), jamming (as seen in the saturation but no decline in rates with adsorption), crystallinity change, cellulase deactivation as a first order process (stochastic modeling studies), and product inhibition (Bommarius et al., 2008).

Based on findings from computational and experimental works, kinetic studies were carried out to identify substrate accessibility and hydrolysability as the major rate hindrances. Reactivity, the rate of hydrolysis per amount of productively bound cellulase, was observed to remain constant up until a conversion level of 66%. Accessibility was quantified by adsorption studies, where the adsorption data at each conversion level was fit to the Langmuir isotherm. While the maximum adsorbable capacity did not decrease much over the conversion range studied (0 – 66%), a steady decrease in the adsorbed amount for various enzyme loadings was observed. This is probably due to a reduction in the cellulase affinity for the cellulose, captured in the association constant of the Langmuir isotherm. Hydrolysability, the reactive fraction of accessible cellulose, was observed to decrease from nearly 100% to approximately 25% at about 30% conversion, and then remain constant. This was also used to determine the fraction of productively bound cellulases at various enzyme loadings, which was shown to follow a trend similar to that of hydrolysability with conversion (at maximum substrate coverage with cellulases, the fraction of productively bounds cellulases is equal to hydrolysability). Enzyme clogging was observed in the form of higher restart rates as compared to the

uninterrupted rates. This is the first work to thoroughly screen the various rate limiting hypotheses, and quantify hydrolysability along conversion.

6.2 Multivariate statistical analysis to determine the degree of crystallinity of cellulose

By quantifying the respective contributions of amorphous and crystalline cellulose to the X-ray diffraction spectra of cellulose with intermediate degrees of crystallinity (Avicel and fibrous cellulose), a new method to obtain consistent crystallinity index values was developed. The crystallinity indices obtained were found to be linearly related to the enzymatic hydrolysis rates. Dimensionality reduction of the spectra with principal component analysis revealed the single dimensional nature of the spectra, and was also used to determine the crystallinity index values. Crystallinity values obtained from regressing the whole spectrum, PCA, and leave-one-out validation overlapped very well with each other. The calculated crystallinity values of cellulose mixtures prepared with varying ratios of Avicel and PASC matched very well with the theoretical values. Prediction of hydrolysis rates with X-ray spectra was also shown to be possible by regressing the hydrolysis rates to the principal component scores and the crystallinity index values.

6.3 Computational analysis of the protein sequence space to identify target mutations

The PCA method developed to identify target mutations exploits the concealed pattern in a protein family's sequences. The sequence of interest is approximated by one that is closer to the identified landscape, and the differences in residues at different positions are chosen as the target mutations. The success in identifying beneficial mutations was demonstrated through the PCA method's application to the family of proteinase K. Effects of mutations at certain positions were compared with published results, and a

major fraction of the top ranked mutations were seen to be positive. Performance of the PCA method on old yellow enzymes was tested by picking the suggested mutations in the first and second shell of the bound flavin and cyclohexenone substrate. When tested for activity and enantioselectivity, no significant improvement over the WT (ene reductase from *Yersinia bercovieri,* Yers-ER) was observed. One reason for the lack of success of PCA on Yers-ER could be that the first and second shell mutations selected from the PCA list were not ranked very high (highest rank was 25). Despite the lack of gain in activity, only a few showed a drastic reduction in activity, pointing to a possible use of the PCA method in identifying tolerable mutations; a control study is however required to confirm this.

6.4 Recommendations for future work

6.4.1 Stochastic modeling and kinetic studies on lignocellulosic substrates and pure cellulases

The Markov model approach used in this thesis to check for clogging of cellobiohydrolases, can be investigated further based on parameters estimated from kinetic studies. Once validated with current time-conversion data, it can be used to check for parameter sensitivity, which can give insights into the properties of cellulases or substrates to engineer to achieve higher rates. Stochastic modeling can also be used to study fractal kinetics by introducing check blocks and obstacles, and relate it to the clogging phenomenon. It should be possible to use the stochastic model to examine synergism by varying the amounts of endoglucanases and cellobiohydrolases.

The kinetic experiments in this work were carried out using a pure cellulosic substrate, and an enzyme mixture. If done with pure cellobiohydrolases, the hydrolysable fraction of the reducing ends can be determined, a quantity not known yet. If it seems laborious and costly (due to limited pure protein availability) to obtain data at different conversion levels (due to desorption and drying procedure), obtaining the data even for

unconverted cellulose will be helpful. The same kinetic studies on Cel7A CBD pretreated cellulose can reveal the real reason for enhancement in rates. These should also be extended to lignocellulosic substrates. The current experimental design is based on the concept of a pure cellulosic substrate consisting of accessible, hydrolysable, and inaccessible portions. If lignin is modeled to have accessible and inaccessible portions, results of kinetic experiments correlated with the fraction of lignin remaining can give insights into the role of lignin in governing the rates.

The rate expression used to design the restart and adsorption experiments was formulated mainly to tease apart the different factors governing the rates, and has limited predictive capability. The parameters in this expression change with time and conversion, and are probably a manifestation of more fundamental phenomena. To model these basic phenomena with constant parameters can be challenging, but without it, a predictive model will not be possible.

6.4.2 Using multivariate statistical analysis on X-ray data for characterization of lignocellulosic substrates

The crystallinity method developed was tested on pure cellulosic substrates by quantifying the contributions of amorphous and crystalline portions. When extending to lignocellulosic substrates, the lignin spectrum will also have to be taken into consideration, thus adding another parameter. Although lignin is known to be almost amorphous, its statistical significance/insignificance in contributing to the substrate's spectrum must be ascertained. If the fraction of lignin in a substrate is known, then the parameter corresponding to lignin's contribution to the spectrum can be fixed; this fraction is however, not know beforehand in many cases. It will also be interesting to compare the crystallinities and the X-ray spectra of the cellulose component only among the different lignocellulosic substrates.

It was shown that if Avicel and fibrous cellulose data sets are subject to PCA together, two principal components are required to explain the variance in the data. Using lignocellulosic substrates, one principal component might be sufficient to explain data from one substrate, but in a mixture of lignocellulosic substrates it might be interesting to see if the number of principal components required is the same as the number of substrates in the mixture. As for the prediction of hydrolysis rates, different linear curves were obtained for Avicel and fibrous cellulose, probably because crystallinity is not the only property governing the rates. Correlating hydrolysis rates of lignocellulosic substrates with crystallinities, and principal component scores can tell us how many parameters are needed to predict the rates.

Another interesting analysis can include correlating crystallinity index values and long residence time rates, as opposed to the initial rates (glucose produced in 2 minutes) used in this thesis work. The initial rates were observed to correlate with only the first principal component scores, but long term digestibility might correlate with scores from more than one principal components.

6.4.2 Principal component analysis and other dimensionality reduction techniques for protein engineering

The PCA method in this work used only the sequence data to elucidate the pattern in them to suggest mutations. This is a case of unsupervised learning. In the case of available activity data (*y* data), techniques other than PCA like NMF and Isomap can be useful. Isomap, though not conducive for reconstruction of the data set, captures the variance in a smaller number of dimensions and might be very useful for regression purposes. One issue while applying any dimensionality reduction method to protein sequences will be the resolution with respect to single point mutations. Because positions of the order of 300 are collapsed on to very few dimensions (of the order of 1 to 10), for a classification or regression purposes, resolution might be an issue when comparing

variants to the entire family. This is one of the reasons why PCA is so strong; with reconstruction, it is able to suggest those small steps that can be taken towards the landscape. However, if the space of mutations (positions and residues) is defined, then it may be possible to achieve the necessary resolution. In the presence of a high throughput assay, these methods can be used to perform regression tasks similar to those used in ProSAR (Fox et al., 2007).

It might also be interesting to accurately quantify the covariations of residues at certain positions. Principal component analysis implicitly accounts for the covariations, but it must be emphasized that gauging insight into correlations based on PCA is not obvious. The reason is that the quantity of interest is the likelihood of occurrence of two mutations when the WT is mapped back to the sequence space, and is not simply the co-occurrence probability of two residues at given positions in the family of sequences. Recently a few works on analyzing residue coupling with graphical models have been published (Thomas et al., 2008; Thomas et al., 2009), and can be explored further to link with PCA.

One of the properties of the current PCA method is that the suggested residue is within the library of sequences. This is due to the feature space selected. However, if the feature space is selected such that each position is assigned only one dimension with an entry corresponding to a chosen physiological property, then reconstruction can suggest a residue outside the library. The choice of the physiological property is up to the user, and mutations suggested from different physiological properties can be compared.

APPENDIX A

EFFECTS OF DESORPTION PROCEDURE ON AVICEL

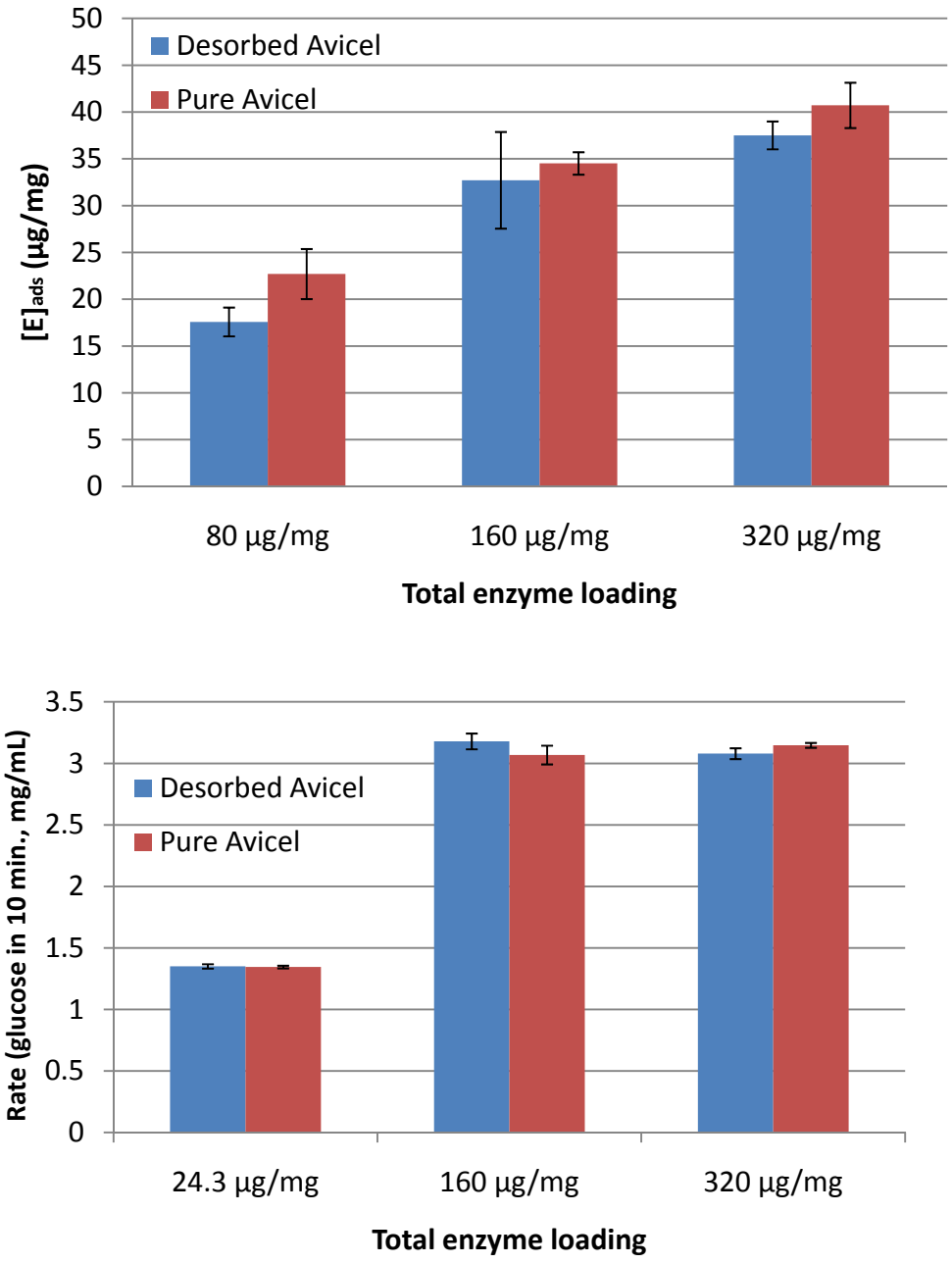


Figure 30. Effects of the desorption procedure on adsorption and hydrolysis on Avicel.

APPENDIX B

EXPERIMENTAL PROCEDURE OF PURIFICATION OF CEL7A AND DETERMINATION OF CHAIN ENDS PER AMOUNT OF SUBSTRATE

Cel7A purification

Cel7A was purified (either from *Trichoderma reesei* expression medium or from commercial cellulase cocktail) by means of anion-exchange chromatography as previously published (Hall et al., 2010a). Purity was confirmed by SDS-PAGE (Hall et al., 2011). After purification, Cel7A buffer was exchanged to sodium acetate buffer (50 mM, pH 5) using a polyethersulfone membrane (molecular weight cut-off of 10 kDa) in a Macrosep device.

Determination of chain ends per amount of substrate

The number of chain ends per amount of cellulose was determined as previously published (Zhang and Lynd, 2005) by measuring the reducing end concentration.

APPENDIX C

EXPERIMENTAL PROCEDURES FOR CHAPTER 4

Material and chemicals

All chemicals and reagents were purchased from Sigma-Aldrich-Fluka unless otherwise stated. Avicel PH-101, mildly acid-washed birch wood (Fluka 11363), fibrous cellulose from cotton linters (Sigma C6288, medium), cellulases from *Trichoderma reesei* and β-glucosidase (from almonds, 5.2 U/mg) were from Sigma, phosphoric acid (85%) was from EMD (Gibbstown, NJ).

Phosphoric acid pretreatment

1 g of slightly moistened Avicel was added to 30 ml of an ice-cold aqueous phosphoric acid solution (concentration ranging from 42% up to 85% wt) and allowed to react over 40 min with occasional stirring. After addition of 20 ml of ice-cold acetone and subsequent stirring, the resulting slurry was filtered over a fritted filtered-funnel and washed three times with 20 ml ice-cold acetone, and 4 times with 100 ml water. The resulting cellulose was used as such in enzymatic hydrolysis experiments, after moisture content was estimated upon oven-drying at 60°C overnight. Samples were freeze-dried prior to X-ray diffraction measurements. The same procedure was followed with fibrous cellulose.

Enzymatic hydrolysis of cellulose

A suspension of cellulose (20 g/l) in sodium acetate buffer (1 ml, 50 mM, pH 5) was hydrated during 1 h under stirring at 50°C. β-Glucosidase (15 U/ml) and cellulases (1.10 mg/ml total protein, 3.8 FPU/ml) were added and the mixture was stirred at 50°C for 2 min. Samples were centrifuged, and glucose content in the supernatant was measured *via*

the dinitrosalicylic acid (DNS) assay. All the samples were run in duplicate and for each of these samples, duplicate assay reading were acquired as well. The mean values are reported as no significant deviation was observed.

Determination of glucose content

Glucose released from cellulose was measured using the DNS assay, as published before (Bommarius et al., 2008). The calibration curve was generated with pure glucose standards.

X-ray diffraction

X-ray diffraction patterns of cellulose samples obtained after freeze-drying were recorded with an X'Pert PRO X-ray diffractometer at room temperature from 10-60 °, using Cu/ $K\alpha_1$ irradiation (1.54 Å) at 45 kV and 40 mA. Scan speed was 0.021425 °/sec with a step size of 0.0167 °.

Solid state ^{13}C NMR

The solid-state cross polarization/magic angle spinning (CP/MAS) ^{13}C-NMR experiments were performed on a Bruker Avance/DSX-400 spectrometer operating at frequencies of 100.55 MHz for ^{13}C. All the experiments were carried out at ambient temperature using a Bruker 4-mm MAS probe. The samples (~ 35% moisture content) were packed in 4 mm zirconium dioxide rotors and spun at 10 kHz. Acquisition was carried out with a CP pulse sequence using 5 μs pulse and 2.0 ms contact pulse. CrI was calculated according to literature (Bommarius et al., 2008).

APPENDIX D

SUPPLEMENTARY FIGURES AND TABLE FOR CHAPTER 4

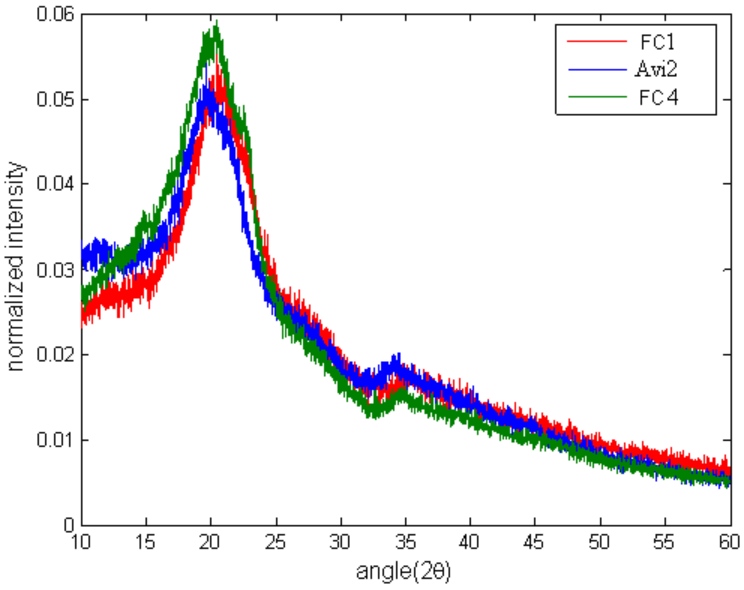


Figure 31. Superimposed spectra of amorphous samples of Avicel (Avi2 - 82.37% acid-pretreated, blue spectrum) and FC (FC1 - 85.00% acid-pretreated, red spectrum). For comparison FC pretreated with 81.71% phosphoric acid is also shown (FC4, green spectrum).

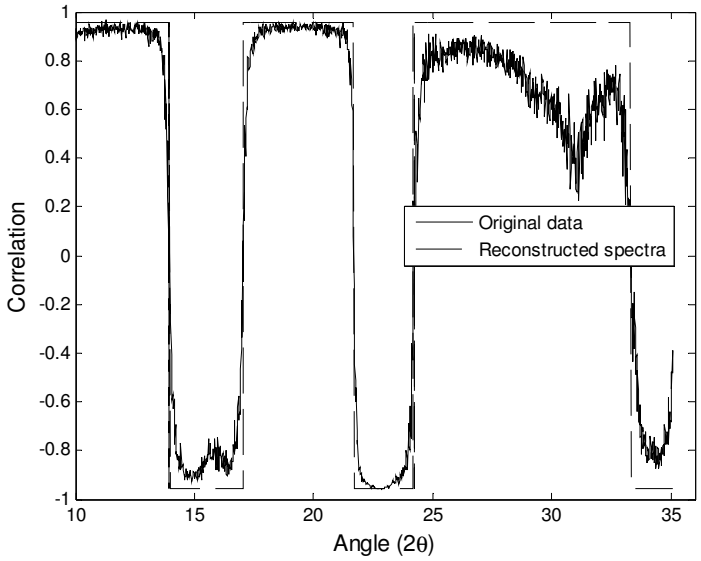


Figure 32. Correlation of the hydrolysis rates with intensities at different diffraction angles for the original spectra and the spectra reconstructed from PCA for FC.

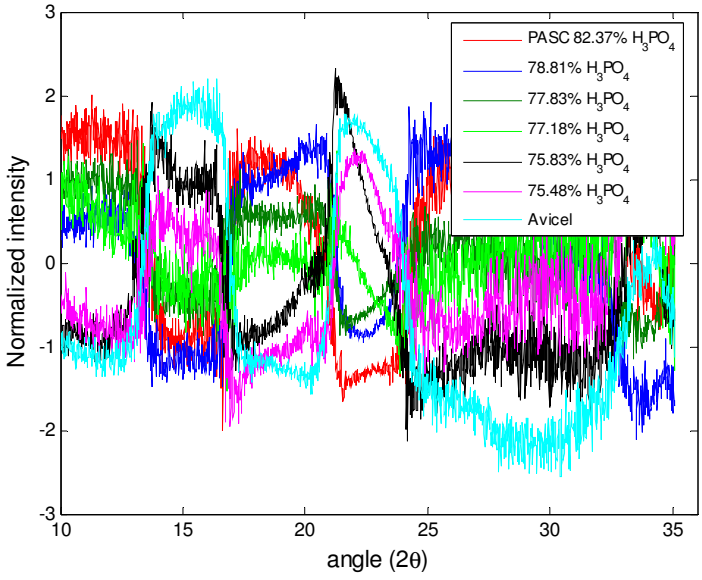


Figure 33. The Z matrix values of untreated and phosphoric acid treated Avicel (when divided by the standard deviation) plotted vs. the diffraction angles.

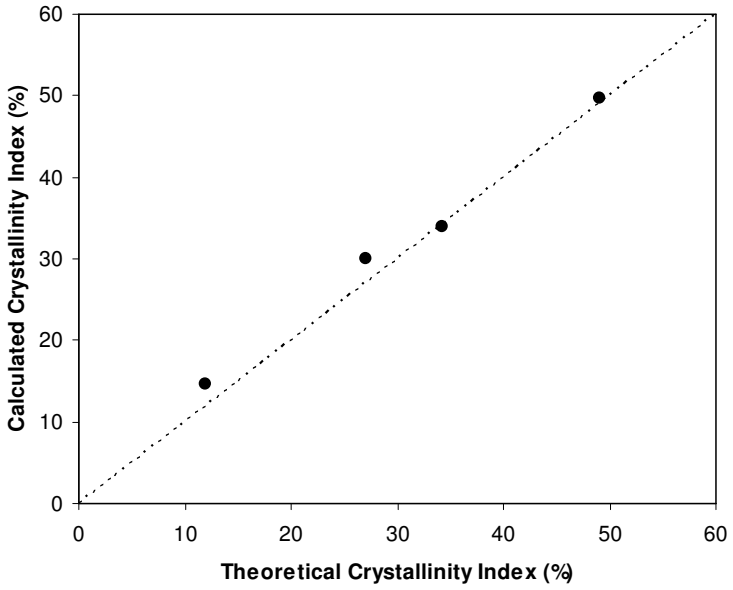


Figure 34. Calculated crystallinity index vs. theoretical crystallinity index for samples prepared by mixing Avicel and amorphous cellulose. Theoretical Cri = (Avicel fraction*Cri_C + amorphous fraction*5), Cri_C = 60% (The cellulose sample used for preparing the samples was found to be not completely amorphous and had a calculated Cri of 5%). The broken line is the y = x line.

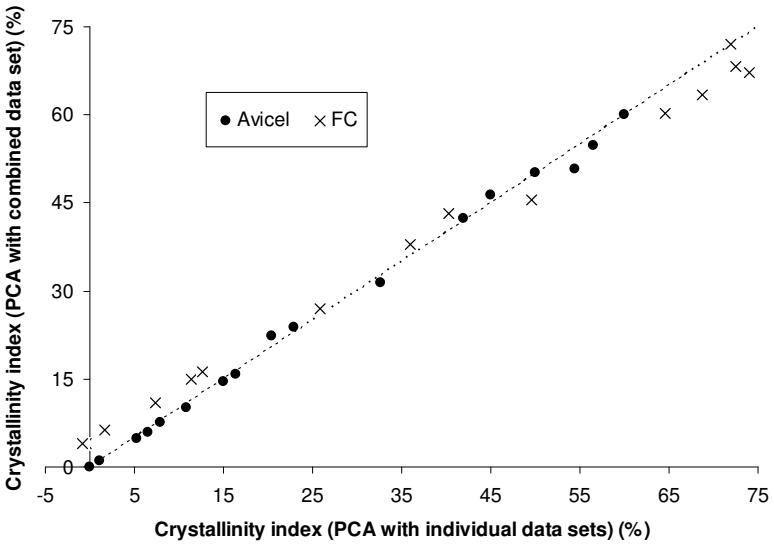


Figure 35. Plot of crystallinity index (%) as calculated with PCA on the combined data set vs. crystallinity index (%) as calculated with PCA on the individual data sets, for Avicel and FC. The broken line is the y = x line.

Table 14. R^2 values (1: fit between the spectra from equation 9 and the original spectra, 2: fit between the spectra reconstructed from one PC and the original spectra) for a) Avicel and b) FC.

a)

1	2	3	4	5
Sample name	**Acid concentration (%)**	**Hydrolysis rate (mg/ml)**	**R^2 (1)**	**R^2 (2)**
Avi1	82.41	8.74	0.93	0.98
Avi2 (PASC)	82.37	9.23	1.00	0.95
Avi3	79.64	9.27	0.96	0.96
Avi4	79.23	7.9	0.91	0.97
Avi5	78.81	7.85	0.93	0.98
Avi6	78.6	7.1	0.90	0.96
Avi7	78.35	6.3	0.94	0.96
Avi8	77.83	6.37	0.98	0.99
Avi9	77.18	4.86	0.97	0.97
Avi10	76.79	4.76	0.91	0.90
Avi11	76.49	4.2	0.96	0.96
Avi12	76.12	3.55	0.99	0.99
Avi13	75.83	2.98	0.97	0.98
Avi14	75.48	2.07	0.98	0.98
Avi15	70.81	2.05	0.91	0.91
Avi16	41.56	1.79	0.95	0.95
Avicel	0	1.2	1.00	1.00

b)

1	2	3	4	5
Sample name	**Acid concentration (%)**	**Hydrolysis rate (mg/ml)**	**R^2 (1)**	**R^2 (2)**
FC2	82.03	9.16	0.94	0.96
FC3	81.78	9.36	0.96	0.99
FC4	81.71	9.74	0.91	0.97
FC5	81.5	6.74	0.92	0.97
FC6	81.22	10.29	0.98	0.99
FC7	81.06	8.7	0.98	0.99
FC8	81.05	3.55	0.97	0.99
FC9	80.71	7.97	0.97	0.96
FC10	80.46	5.63	0.98	0.99
FC11	79.94	5.02	0.98	0.99
FC12	79.51	2.72	0.97	1.00
FC13	78.09	1.61	0.98	1.00
FC14	75.07	1.22	0.97	0.99
FC15	65.22	0.77	0.96	0.99
FC	0	0.57	1.00	0.98

APPENDIX E

PROOF OF PRESERVATION OF CONSERVED AND PRECLUSION OF ABSENT RESIDUES AT A POSITION WITH PCA RECONSTRUCTION

Any mean centered data matrix Z_{pxn} can be expressed as -:

$$Z = U_{pxp}S_{pxn}V^{T}_{nxn} = \sum_{k=1}^{n} \mathbf{u}_k \sigma_k \mathbf{v}^{T}_{k} \quad (1)$$

$\mathbf{u}_k$'s, and $\mathbf{v}_k$'s are the orthonormal vectors of U and V respectively.

The reconstructed matrix Z_r is computed by summing the above expression upto the number of principal components chosen for reconstruction.

$$Z_r = \sum_{k=1}^{\mathbf{Npc}} \mathbf{u}_k \sigma_k \mathbf{v}^{T}_{k} \quad (2)$$

Therefore, the ij^{th} element of Z_r is given by -:

$$(Z_r)_{ij} = \sum_{k=1}^{\mathbf{Npc}} \mathbf{u}_{ki} \sigma_k \mathbf{v}^{T}_{kj} \quad (3)$$

If the i^{th} element of all the $\mathbf{u}$'s are zero, then the i^{th} element of all the columns in Z_r will also be zero. From the original SVD decomposition,

$$Z = U_{pxp}S_{pxn}V^{T}_{nxn} = \sum_{k=1}^{n} \mathbf{u}_k \sigma_k \mathbf{v}^{T}_{k} \quad (4)$$

Post multiplying by $\mathbf{v}^{T}_{k}$ (and ultilizing the orthonormal property), $\mathbf{u}_k$ is given by:

$$\mathbf{u}_k = Z\mathbf{v}_k/\sigma_k \quad (5)$$

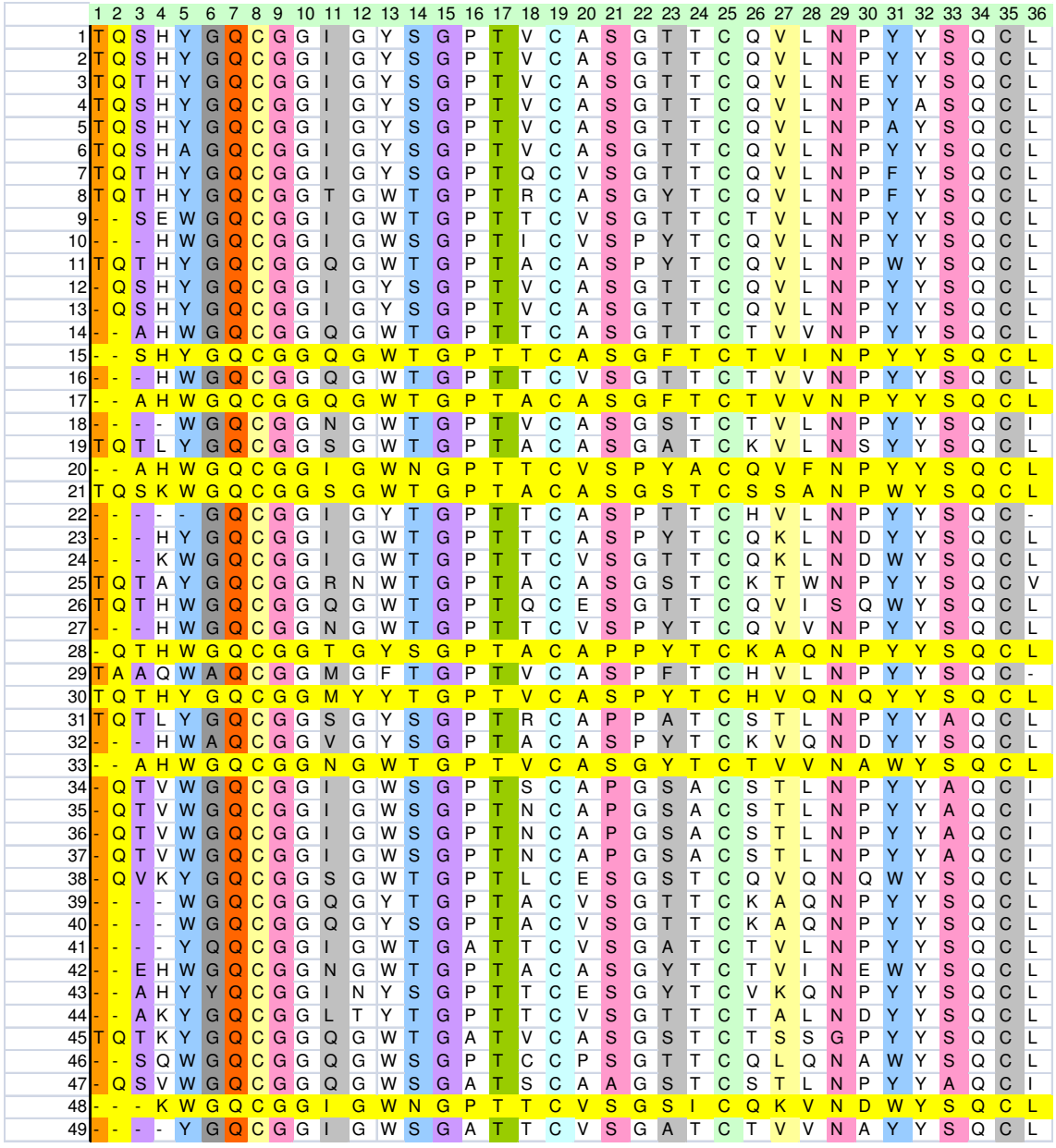
$$\mathbf{u}_{ki} = \sum_{j=1}^{n} \mathbf{Z}_{ij} \mathbf{v}_{kj} \quad (6)$$

In the case of completely conserved residues or complete absence of a residue at any given position Z_{ij} will equal zero, so all the corresponding elements in the $\mathbf{u}$ vectors will also be zero, thereby having no effect on the reconstruction.

APPENDIX F

SEQUENCES FROM FAMILY I OF CELLULOSE BINDING DOMAINS AND THEIR SOURCE

The color scheme is chosen simply to make it easy to follow a position.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	T	Q	S	H	Y	G	Q	C	G	G	I	G	Y	S	G	P	T	V	C	A	S	G	T	T	C	Q	V	L	N	P	Y	Y	S	Q	C	L
2	T	Q	S	H	Y	G	Q	C	G	G	I	G	Y	S	G	P	T	V	C	A	S	G	T	T	C	Q	V	L	N	P	Y	Y	S	Q	C	L
3	T	Q	T	H	Y	G	Q	C	G	G	I	G	Y	S	G	P	T	V	C	A	S	G	T	T	C	Q	V	L	N	E	Y	Y	S	Q	C	L
4	T	Q	S	H	Y	G	Q	C	G	G	I	G	Y	S	G	P	T	V	C	A	S	G	T	T	C	Q	V	L	N	P	Y	A	S	Q	C	L
5	T	Q	S	H	Y	G	Q	C	G	G	I	G	Y	S	G	P	T	V	C	A	S	G	T	T	C	Q	V	L	N	P	A	Y	S	Q	C	L
6	T	Q	S	H	A	G	Q	C	G	G	I	G	Y	S	G	P	T	V	C	A	S	G	T	T	C	Q	V	L	N	P	Y	Y	S	Q	C	L
7	T	Q	T	H	Y	G	Q	C	G	G	I	G	Y	S	G	P	T	Q	C	V	S	G	T	T	C	Q	V	L	N	P	F	Y	S	Q	C	L
8	T	Q	T	H	Y	G	Q	C	G	G	T	G	W	T	G	P	T	R	C	A	S	G	Y	T	C	Q	V	L	N	P	F	Y	S	Q	C	L
9	-	-	S	E	W	G	Q	C	G	G	I	G	W	T	G	P	T	T	C	V	S	G	T	T	C	T	V	L	N	P	Y	Y	S	Q	C	L
10	-	-	-	H	W	G	Q	C	G	G	I	G	W	S	G	P	T	I	C	V	S	P	Y	T	C	Q	V	L	N	P	Y	Y	S	Q	C	L
11	T	Q	T	H	Y	G	Q	C	G	G	Q	G	W	T	G	P	T	A	C	A	S	P	Y	T	C	Q	V	L	N	P	W	Y	S	Q	C	L
12	-	Q	S	H	Y	G	Q	C	G	G	I	G	Y	S	G	P	T	V	C	A	S	G	T	T	C	Q	V	L	N	P	Y	Y	S	Q	C	L
13	-	Q	S	H	Y	G	Q	C	G	G	I	G	Y	S	G	P	T	V	C	A	S	G	T	T	C	Q	V	L	N	P	Y	Y	S	Q	C	L
14	-	-	A	H	W	G	Q	C	G	G	Q	G	W	T	G	P	T	T	C	A	S	G	T	T	C	T	V	V	N	P	Y	Y	S	Q	C	L
15	-	-	S	H	Y	G	Q	C	G	G	Q	G	W	T	G	P	T	T	C	A	S	G	F	T	C	T	V	I	N	P	Y	Y	S	Q	C	L
16	-	-	-	H	W	G	Q	C	G	G	Q	G	W	T	G	P	T	T	C	V	S	G	T	T	C	T	V	V	N	P	Y	Y	S	Q	C	L
17	-	-	A	H	W	G	Q	C	G	G	Q	G	W	T	G	P	T	A	C	A	S	G	F	T	C	T	V	V	N	P	Y	Y	S	Q	C	L
18	-	-	-	-	W	G	Q	C	G	G	N	G	W	T	G	P	T	V	C	A	S	G	S	T	C	T	V	L	N	P	Y	Y	S	Q	C	I
19	T	Q	T	L	Y	G	Q	C	G	G	S	G	W	T	G	P	T	A	C	A	S	G	A	T	C	K	V	L	N	S	Y	Y	S	Q	C	L
20	-	-	A	H	W	G	Q	C	G	G	I	G	W	N	G	P	T	T	C	V	S	P	Y	A	C	Q	V	F	N	P	Y	Y	S	Q	C	L
21	T	Q	S	K	W	G	Q	C	G	G	S	G	W	T	G	P	T	A	C	A	S	G	S	T	C	S	S	A	N	P	W	Y	S	Q	C	L
22	-	-	-	-	-	G	Q	C	G	G	I	G	Y	T	G	P	T	T	C	A	S	P	T	T	C	H	V	L	N	P	Y	Y	S	Q	C	-
23	-	-	-	H	Y	G	Q	C	G	G	I	G	W	T	G	P	T	T	C	A	S	P	Y	T	C	Q	K	L	N	D	Y	Y	S	Q	C	L
24	-	-	-	K	W	G	Q	C	G	G	I	G	W	T	G	P	T	T	C	V	S	G	T	T	C	Q	K	L	N	D	W	Y	S	Q	C	L
25	T	Q	T	A	Y	G	Q	C	G	G	R	N	W	T	G	P	T	A	C	A	S	G	S	T	C	K	T	W	N	P	Y	Y	S	Q	C	V
26	T	Q	T	H	W	G	Q	C	G	G	Q	G	W	T	G	P	T	Q	C	E	S	G	T	T	C	Q	V	I	S	Q	W	Y	S	Q	C	L
27	-	-	-	H	W	G	Q	C	G	G	N	G	W	T	G	P	T	T	C	V	S	P	Y	T	C	Q	V	V	N	P	Y	Y	S	Q	C	L
28	-	Q	T	H	W	G	Q	C	G	G	T	G	Y	S	G	P	T	A	C	A	P	P	Y	T	C	K	A	Q	N	P	Y	Y	S	Q	C	L
29	T	A	A	Q	W	A	Q	C	G	G	M	G	F	T	G	P	T	V	C	A	S	P	F	T	C	H	V	L	N	P	Y	Y	S	Q	C	-
30	T	Q	T	H	Y	G	Q	C	G	G	M	Y	Y	T	G	P	T	V	C	A	S	P	Y	T	C	H	V	Q	N	Q	Y	Y	S	Q	C	L
31	T	Q	T	L	Y	G	Q	C	G	G	S	G	Y	S	G	P	T	R	C	A	P	P	A	T	C	S	T	L	N	P	Y	Y	A	Q	C	L
32	-	-	-	H	W	A	Q	C	G	G	V	G	Y	S	G	P	T	A	C	A	S	P	Y	T	C	K	V	Q	N	D	Y	Y	S	Q	C	L
33	-	-	A	H	W	G	Q	C	G	G	N	G	W	T	G	P	T	V	C	A	S	G	Y	T	C	T	V	V	N	A	W	Y	S	Q	C	L
34	-	Q	T	V	W	G	Q	C	G	G	I	G	W	S	G	P	T	S	C	A	P	G	S	A	C	S	T	L	N	P	Y	Y	A	Q	C	I
35	-	Q	T	V	W	G	Q	C	G	G	I	G	W	S	G	P	T	N	C	A	P	G	S	A	C	S	T	L	N	P	Y	Y	A	Q	C	I
36	-	Q	T	V	W	G	Q	C	G	G	I	G	W	S	G	P	T	N	C	A	P	G	S	A	C	S	T	L	N	P	Y	Y	A	Q	C	I
37	-	Q	T	V	W	G	Q	C	G	G	I	G	W	S	G	P	T	N	C	A	P	G	S	A	C	S	T	L	N	P	Y	Y	A	Q	C	I
38	-	Q	V	K	Y	G	Q	C	G	G	S	G	W	T	G	P	T	L	C	E	S	G	S	T	C	Q	V	Q	N	Q	W	Y	S	Q	C	L
39	-	-	-	-	W	G	Q	C	G	G	Q	G	Y	T	G	P	T	A	C	V	S	G	T	T	C	K	A	Q	N	P	Y	Y	S	Q	C	L
40	-	-	-	-	W	G	Q	C	G	G	Q	G	Y	S	G	P	T	A	C	V	S	G	T	T	C	K	A	Q	N	P	Y	Y	S	Q	C	L
41	-	-	-	-	Y	Q	Q	C	G	G	I	G	W	T	G	A	T	T	C	V	S	G	A	T	C	T	V	L	N	P	Y	Y	S	Q	C	L
42	-	-	E	H	W	G	Q	C	G	G	N	G	W	T	G	P	T	A	C	A	S	G	Y	T	C	T	V	I	N	E	W	Y	S	Q	C	L
43	-	-	A	H	Y	Y	Q	C	G	G	I	N	Y	S	G	P	T	T	C	E	S	G	Y	T	C	V	K	Q	N	P	Y	Y	S	Q	C	L
44	-	-	A	K	Y	G	Q	C	G	G	L	T	Y	T	G	P	T	T	C	V	S	G	T	T	C	T	A	L	N	D	Y	Y	S	Q	C	L
45	T	Q	T	K	Y	G	Q	C	G	G	Q	G	W	T	G	A	T	V	C	A	S	G	S	T	C	T	S	S	G	P	Y	Y	S	Q	C	L
46	-	-	S	Q	W	G	Q	C	G	G	Q	G	W	S	G	P	T	C	C	P	S	G	T	T	C	Q	L	Q	N	A	W	Y	S	Q	C	L
47	-	Q	S	V	W	G	Q	C	G	G	Q	G	W	S	G	A	T	S	C	A	A	G	S	T	C	S	T	L	N	P	Y	Y	A	Q	C	I
48	-	-	-	K	W	G	Q	C	G	G	I	G	W	N	G	P	T	T	C	V	S	G	S	I	C	Q	K	V	N	D	W	Y	S	Q	C	L
49	-	-	-	-	Y	G	Q	C	G	G	I	G	W	S	G	A	T	T	C	V	S	G	A	T	C	T	V	V	N	A	Y	Y	S	Q	C	L

Sequences 15, 17, 20, 21, 28, 30, 33, and 48 were excluded from PCA analysis.

Source:

No.	Enzyme
1	T reesei CBH1
2	T. viride
3	Trichoderma sp. XST1 CBH1
4	T reesei CBH1 mutated
5	T reesei CBH1 mutated
6	T reesei CBH1 mutated
7	Hypocrea virens
8	Hypocrea lixii CBH
9	P Chrysoporium exocbh
10	endo-1,4-xylanase D [Penicillium funiculosum]
11	endo-1,4-xylanase D Penicillium funiculosum
12	endoglucanase [Aspergillus fumigatus Af293]
13	unnamed protein product [Sordaria macrospora]
14	xylanase/cellobiohydrolase [Penicillium funiculosum]
15	1,4-beta-D-glucan-cellobiohydrolyase, putative [Talaromyces stipitatus]
16	cellobiohydrolase I [Penicillium occitanis]
17	1,4-beta-D-glucan-cellobiohydrolyase, putative [Penicillium marneffei]
18	unnamed protein product [Podospora anserina]
19	glycosyl hydrolase family 45 protein Neosartorya fischeri]
20	endo-1,4-beta-xylanase, putative [Penicillium marneffei]
21	xylosidase/glycosyl hydrolase, putative [Neosartorya fischeri]
22	CBHI [Volvariella volvacea]
23	cellobiohydrolyase [Aspergillus fumigatus]
24	endo-1,4-beta-xylanase [Aspergillus fumigatus]
25	Glycosyl hydrolase family 61 [Neosartorya fischeri]
26	acetyl xylan esterase [Hypocrea jecorina]
27	endo-1,4-beta-xylanase, putative [Talaromyces stipitatus]
28	endoglucanase, putative [Penicillium marneffei]
29	cellulase [Irpex lacteus]
30	endoglucanase, putative [Talaromyces stipitatus]
31	endoglucanase IV [Hypocrea jecorina]
32	endoglucanase 1 [Penicillium echinulatum]
33	acetyl xylan esterase, putative [Neosartorya fischeri]
34	endoglucanase II [Trichoderma viride]
35	endoglucanase III [Trichoderma viride]
36	endoglucanase III [Hypocrea jecorina]
37	endoglucanase II [Hypocrea jecorina]
38	glycoside hydrolase family 5 [Nectria haematococca]
39	endoglucanase I [Penicillium oxalicum]
40	endoglucanase I [Penicillium decumbens]
41	cellulose-binding beta-glucosidase [Phanerochaete chrysosporium]
42	acetyl xylan esterase [Aspergillus fumigatus]
43	endoglucanase/cellulase, putative [Aspergillus flavus]
44	family 61 endoglucanase [Phanerochaete chrysosporium]
45	cellulase CEL7A [Lentinula edodes]
46	exoglucanase [Verticillium albo-atrum]
47	cellobiohydrolase II [Acremonium cellulolyticus]
48	endo-1,4-beta-xylanase, putative [Neosartorya fischeri]
49	cellulase CEL6B [Lentinula edodes]

REFERENCES

Aita, T., Iwakura, M., Husimi, Y. 2001. A cross-section of the fitness landscape of dihydrofolate reductase. *Protein Eng.*, **14**(9), 633-638.

Aita, T., Uchiyama, H., Inaoka, T., Nakajima, M., Kokubo, T., Husimi, Y. 2000. Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: Application to prolyl endopeptidase and thermolysin. *Biopolymers*, **54**(1), 64-79.

Al-Zuhair, S. 2008. The effect of crystallinity of cellulose on the rate of reducing sugars production by heterogeneous enzymatic hydrolysis. *Bioresour. Technol.*, **99**, 4078-4085.

Anacker, L.W., Kopelman, R. 1987. Steady-state Chemical-kinetics on Fractals - Segregation of Reactants. *Phys. Rev. Lett.*, **58**(4), 289-291.

Ardizzone, S., Dioguard, F.S., Mussini, T., Mussini, P.R., Rondinini, S., Vercelli, B., Vertova, A. 1999. Microcrystalline cellulose powders: structure, surface features and water sorption capability. *Cellulose*, **6**, 57-69.

Arnold, F.H., Volkov, A.A. 1999. Directed evolution of biocatalysts. *Curr. Opin. Chem. Biol.*, **3**(1), 54-59.

Asenjo, J.A. 1983. Maximizing the Formation of Glucose in the Enzymatic Hydrolysis of Insoluble Cellulose. *Biotechnol. Bioeng.*, **25**, 3185-3190.

Asenjo, J.A. 1984. Modelling The Bioconversion of Cellulose Into Microbial Products: Rate Limitations. *Process Biochem.*, **19**, 217-224.

Bader, J., Bellgardt, K.-H., Singh, A., Kumar, P.K.R., Schugerl, K. 1992. Modelling and simulation of cellulase adsorption and recycling during enzymatic hydrolysis of cellulosic materials. *Bioprocess Eng.*, **7**, 235-240.

Bansal, P., Hall, M., Realff, M.J., Lee, J.H., Bommarius, A.S. 2009. Modeling Cellulase Kinetics On Lignocellulosic Substrates. *Biotechnol. Adv.*, **27**, 833-848.

Bansal, P., Hall, M., Realff, M.J., Lee, J.H., Bommarius, A.S. 2010. Multivariate statistical analysis of X-ray data from cellulose: A new method to determine degree of crystallinity and predict hydrolysis rates. *Bioresour. Technol.*, **101**(12), 4461-4471.

Barak, Y., Nov, Y., Ackerley, D.F., Matin, A. 2008. Enzyme improvement in the absence of structural knowledge: a novel statistical approach. *Isme Journal*, **2**(2), 171-179.

Barnette, A.L., Bradley, L.C., Veres, B.D., Schreiner, E.P., Park, Y.B., Park, J., Park, S., Kim, S.H. 2011. Selective Detection of Crystalline Cellulose in Plant Cell Walls with Sum-Frequency-Generation (SFG) Vibration Spectroscopy. *Biomacromolecules*, **12**, 2434–2439.

Beckham, G.T., Matthews, J.F., Peters, B., Bomble, Y.J., Himmel, M.E., Crowley, M.F. 2011. Molecular-Level Origins of Biomass Recalcitrance: Decrystallization Free Energies for Four Common Cellulose Polymorphs. *J. Phys. Chem. B*, **115**(14), 4118-4127.

Beldman, G., Voragen, A.G.J., Rombouts, F.M., Leeuwen, M.F.S.-v., Pilnik, W. 1987. Adsorption and Kinetic Behaviour of Purified Endoglucanases and Exoglucanases from *Trichoderma viride*. *Biotechnol. Bioeng.*, **30**, 251-257.

Beldman, G., Voragen, A.G.J., Rombouts, F.M., Pilnik, W. 1988. Synergism in Cellulose Hydrolysis by Endoglucanases and Exoglucanases Purified from Trichoderma viride. *Biotechnol. Bioeng.*, **31**, 173-178.

Beltrame, P.L., Carniti, P., Focher, B., Marzetti, A., Sarto, V. 1984. Enzymatic Hydrolysis of Cellulosic Materials: A Kinetic Study. *Biotechnol. Bioeng.*, **26**, 1233-1238.

Berlin, A., Gilkes, N., Kurabi, A., Bura, R., Tu, M., Kilburn, D., Saddler, J. 2005. Weak Lignin-Binding Enzymes. *Appl. Biochem. Biotech.*, **121-124**, 163-170.

Berlin, A., Maximenko, V., Gilkes, N., Saddler, J. 2007. Optimization of Enzyme Complexes for Lignocellulose Hydrolysis. *Biotechnol. Bioeng.*, **97**(2), 287-296.

Berry, H. 2002. Monte Carlo simulations of enzyme reactions in two dimensions: Fractal kinetics and spatial segregation. *Biophys. J.*, **83**(4), 1891-1901.

Bertran, M.S., Dale, B.E. 1985. Enzymatic Hydrolysis And Recrystallization Behaviour of Initially Amorphous Cellulose. *Biotechnol. Bioeng.*, **27**, 177-181.

Bezerra, R.M.F., Dias, A.A. 2004. Discrimination Among Eight Modified Michaelis-Menten Kinetics Models of Cellulose Hydrolysis With a Large Range of Substrate/Enzyme Ratios. *Appl. Biochem. Biotech.*, **112**, 173-184.

Bezerra, R.M.F., Dias, A.A. 2005. Enzymatic Kinetic of Cellulose Hydrolysis. *Appl. Environ. Microbiol.*, **126**, 49-59.

Bommarius, A.S., Blum, J.K., Abrahamson, M.J. 2011. Status of protein engineering for biocatalysts: how to design an industrially useful biocatalyst. *Curr. Opin. Chem. Biol.*, **15**(2), 194-200.

Bommarius, A.S., Katona, A., Cheben, S.E., Patel, A.S., Ragauskas, A.J., Knudson, K., Pu, Y. 2008. Cellulase kinetics as a function of cellulose pretreatment. *Metab. Eng.*, **10**(6), 370-381.

Borjesson, J., Peterson, R., Tjerneld, F. 2007. Enhanced enzymatic conversion of softwood lignocellulose by poly(ethylene glycol) addition. *Enzyme Microb. Technol.*, **40**(4), 754-762.

Brouk, M., Nov, Y., Fishman, A. 2010. Improving Biocatalyst Performance by Integrating Statistical Methods into Protein Engineering. *Appl. Environ. Microbiol.*, **76**(19), 6397-6403.

Brown, R.E., Jarvis, K.L., Hyland, K.J. 1989. Protein Measurement Using Bicinchonic Acid: Elimination of Interfering Substances. *Anal. Biochem.*, **180**, 136-139.

Brown, R.F., Holtzapple, M.T. 1990. A Comparison of the Michaelis-Menten and HCH-1 Models. *Biotechnol. Bioeng.*, **36**, 1151-1154.

Caminal, G., López-Santín, J., Solà, C. 1985. Kinetic Modeling of the Enzymatic Hydrolysis of Pretreated Cellulose. *Biotechnol. Bioeng.*, **27**, 1282-1280.

Carrard, G., Linder, M. 1999. Widely different off rates of two closely related cellulose-binding domains from Trichoderma reesei. *Eur. J. Biochem.*, **262**(3), 637-643.

Cateto, C., Hu, G., Ragauskas, A. 2011. Enzymatic hydrolysis of organosolv Kanlow switchgrass and its impact on cellulose crystallinity and degree of polymerization *Energy Environ. Sci.*, **4**, 1516-1521.

Chang, V.S., Holtzapple, M.T. 2000. Fundamental Factors Affecting Biomass Enzymatic Reactivity. *Appl. Biochem. Biotech.*, **84-86**, 5-37.

Chen, Y., Stipanovic, A.J., Winter, W.T., Wilson, D.B., Kim, Y.-J. 2007. Effect of digestion by pure cellulases on crystallinity and average chain length for bacterial and microcrystalline celluloses. *Cellulose*, **14**, 283-293.

Chung, F.H., Scott, R.W. 1973. A New Approach to the Determination of Crystallinity of Polymers by X-ray Diffraction. *J. Appl. Crystallogr.*, **6**, 225-230.

Clarkin, S.D., Clesceri, L.S. 2002. Enzymatic hydrolysis and physical characterization of commercial celluloses and cellulose-based ion exchange powdered mixed resin. *Appl. Microbiol. Biotechnol.*, **60**, 485-488.

Converse, A.O., Matsuno, R., Tanaka, M., Taniguchi, M. 1988. A Model of Enzyme Adsorption and Hydrolysis of Microcrystalline Cellulose with Slow Deactivation of the Adsorbed Enzyme. *Biotechnol. Bioeng.*, **32**, 38-45.

Converse, A.O., Ooshima, H., Burns, D.S. 1990. Kinetics of Enzymatic Hydrolysis of Lignocellulosic Materials Based on Surface Area of Cellulose Accessible to Enzyme and Enzyme Adsorption on Lignin and Cellulose. *Appl. Biochem. Biotech.*, **24/25**, 67 - 73.

Converse, A.O., Optekar, J.D. 1993. A Synergistic Kinetics Model for Enzymatic Cellulose Hydrolysis Compared to Degree-of-Synergism Experimental Results. *Biotechnol. Bioeng.*, **42**, 145-148.

Crameri, A., Raillard, S.A., Bermudez, E., Stemmer, W.P.C. 1998. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, **391**(6664), 288-291.

DeanIII, S.W., Rollings, J.E. 1992. Analysis and Quantification of a Mixed Exo-Acting and Endo-Acting Polysaccharide Depolymerization System. *Biotechnol. Bioeng.*, **39**, 968-976.

Debye, P. 1915. Zerstreuung von Röntgenstrahlen. *Annalen der Physik*, **46**, 809-823.

Desai, S.G., Converse, A.O. 1997. Substrate Reactivity as a Function of the Extent of Reaction in the Enzymatic Hydrolysis of Lignocellulose. *Biotechnol. Bioeng.*, **56**(6), 650-655.

Ding, H., Xu, F. 2004. Productive Cellulase Adsorption on Cellulose. *ACS Symp. Ser.*, **889**, 154-169.

Divne, C., Stahlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J.K.C., Teeri, T.T., Jones, A. 1994. The three-dimensional structure of the catalytic core of cellobiohydrolase I from *Trichoderma reesei*. *Science*, **265**, 524-528.

Divne, C., Ståhlberg, J., Teeri, T.T., Jones, T.A. 1998. High-resolution Crystal Structures Reveal How a Cellulose Chain is Bound in the 50 Å Long Tunnel of Cellobiohydrolase I from *Trichoderma reesei*. *J. Mol. Biol.*, **275**, 309-325.

Doelker, E., Gurny, R., Schurz, J., Janosi, A., Matin, N. 1987. Degrees of Crystallinity and Polymerization of Modified Cellulose Powders for Direct Tableting. *Powder Technol.*, **52**, 207-213.

Dourado, F., Gama, F.M., Chibowski, E., Mota, M. 1998. Characterization of cellulose surface free energy. *J. Adhes. Sci. Technol.*, **12**(10), 1081-1090.

Drissen, R.E.T., Maas, R.H.W., Maarel, M.J.E.C.V.D., Kabel, M.A., Schols, H.A., Tramper, J., Beeftink, H.H. 2007. A generic model for glucose production from various cellulose sources by a commercial cellulase complex. *Biocatal. Biotransform.*, **25**(6), 419-429.

Dubey, A., Realff, M.J., Lee, J.H., Bommarius, A.S. 2005. Support vector machines for learning to identify the critical positions of a protein. *J. Theor. Biol.*, **234**(3), 351-361.

Eriksson, T., Karlsson, J., Tjerneld, F. 2002. A Model Explaining Declining Rate in Hydrolysis of Lignocellulose Substrates with Cellobiohydrolase I (Cel7A) and Endoglucanase I (Cel7B) of *Trichoderma reesei*. *Appl. Biochem. Biotech.*, **101**, 41-60.

Fan, L.T., Lee, Y.H. 1983. KINETIC-STUDIES OF ENZYMATIC-HYDROLYSIS OF INSOLUBLE CELLULOSE - DERIVATION OF A MECHANISTIC KINETIC-MODEL. *Biotechnol. Bioeng.*, **25**(11), 2707-2733.

Fenske, J.J., Penner, M.H., Bolt, J.P. 1999. A Simple Individual-based Model of Insoluble Polysaccharide Hydrolysis: the Potential for Autosynergism with Dual-activity Glycosidases. *J. Theor. Biol.*, **199**, 113-118.

Fogler, H.S. 2005. *Elements of Chemical Reaction Engineering*. *4th ed*. Prentice Hall.

Fox, R.J., Davis, S.C., Mundorff, E.C., Newman, L.M., Gavrilovic, V., Ma, S.K., Chung, L.M., Ching, C., Tam, S., Muley, S., Grate, J., Gruber, J., Whitman, J.C., Sheldon, R.A., Huisman, G.W. 2007. Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.*, **25**(3), 338-344.

Fox, R.J., Huisman, G.W. 2008. Enzyme optimization: moving from blind evolution to statistical exploration of sequence-function space. *Trends Biotechnol.*, **26**(3), 132-138.

Fujii, M., Murakami, S., Yamada, Y., Ona, T., Nakamura, T. 1981. A Kinetic Equation for Hydrolysis of Polysaccharides by Mixed Exo- and Endoenzyme systems. *Biotechnol. Bioeng.*, **23**, 1393-1398.

Fujii, M., Shimizu, M. 1986. Synergism of Endoenzyme and Exoenzyme on Hydrolysis of Soluble Cellulose Derivatives. *Biotechnol. Bioeng.*, **28**, 878-882.

Galbe, M., Zacchi, G. 2002. A review of the production of ethanol from softwood. *Appl. Microbiol. Biotechnol.*, **59**, 618-628.

Gama, F.M., Teixeira, J.A., Mota, M. 1994. Cellulose Morphology and Enzymatic Reactivity: A Modified Solute Exclusion Technique. *Biotechnol. Bioeng.*, **43**, 381-387.

Gan, Q., Allen, S.J., Taylor, G. 2003. Kinetic dynamics in heterogeneous enzymatic hydrolysis of cellulose: an overview, an experimental study and mathematical modeling. *Process Biochem.*, **38**, 1003-1018.

Garvey, C.J., Parker, I.H., Simon, G.P. 2005. On the interpretation of X-ray diffraction powder patterns in terms of the nanostructure of cellulose I fibres. *Macromol. Chem. Phys.*, **206**, 1568-1575.

Gharpuray, M.M., Lee, Y.-H., Fan, L.T. 1983. Structural Modification of Lignocellulosics by Pretreatments to Enhance Enzymatic Hydrolysis. *Biotechnol. Bioeng.*, **25**, 157-172.

Ghose, T.K., Bisaria, V.S. 1979. Studies on the Mechanism of Enzymatic Hydrolysis of Cellulosic Substances. *Biotechnol. Bioeng.*, **21**, 131-146.

Gregg, D.J., Saddler, J.N. 1996. Factors Affecting Cellulose Hydrolysis and the Potential of Enzyme Recycle to Enhance the Efficiency of an Integrated Wood To Ethanol Process. *Biotechnol. Bioeng.*, **51**(4), 375-383.

Gupta, R., Lee, Y.Y. 2008. Mechanism of cellulase reaction on pure cellulosic substrates. *Biotechnol. Bioeng.*, **102**(6), 1570-1581.

Gusakov, A.V., Salanovich, T.N., Antonov, A.I., Ustinov, B.B., Okunev, O.N., Burlingame, R., Emalfarb, M., Baez, M., Sinitsyn, A.P. 2007. Design of Highly Efficient Cellulase Mixtures for Enzymatic Hydrolysis of Cellulose. *Biotechnol. Bioeng.*, **97**(5), 1028-1038.

Gusakov, A.V., Sinitsyn, A.P., Klyosov, A.A. 1985. Kinetics of the enzymatic hydrolysis of cellulose: 1. A mathematical model for a batch reactor process. *Enzyme Microb. Technol.*, **7**, 346-352.

Hall, M., Bansal, P., Lee, J.H., Realff, M.J., Bommarius, A.S. 2011. Biological pretreatment of cellulose: Enhancing enzymatic hydrolysis rate using cellulose-binding domains from cellulases *Bioresour. Technol.*, **102**(3), 2910-2915.

Hall, M., Bansal, P., Lee, J.H., Realff, M.J., Bommarius, A.S. 2010a. Cellulose crystallinity - a key predictor of the enzymatic hydrolysis rate. *FEBS J.*, **277**(6), 1571-1582.

Hall, M., Rubin, J., Behrens, S.H., Bommarius, A.S. in press. The Cellulose-Binding Domain of Cellobiohydrolase Cel7A from Trichoderma reesei Is Also a Thermostabilizing Domain. *J. Biotechnol.* doi:10.1016/j.jbiotec.2011.07.016

Hall, M., Yanto, Y., Bommarius, A.S. 2010b. Enzymes, Enolate Reductases "Old Yellow Enzyme". *Encyclopedia of Industrial Biotechnology: Bioprocess, Bioseparation, and Cell Technology*.

Harjunpää, V., Teleman, A., Koivula, A., Ruohonen, L., Teeri, T.T., Teleman, O., Drakenberg, T. 1996. Cello-oligosaccharide hydrolysis by cellobiohydrolase II from *Trichoderma reesei*. *Eur. J. Biochem.*, **240**(3), 584-591.

Heinzelman, P., Snow, C.D., Wu, I., Nguyen, C., Villalobos, A., Govindarajan, S., Minshull, J., Arnold, F.H. 2009. A family of thermostable fungal cellulases created by structure-guided recombination. *Proc. Natl. Acad. Sci. U. S. A.*, **106**(14), 5610-5615.

Henikoff, S., Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 10915-10919.

Henrissat, B. 1994. Cellulases and their interaction with cellulose. *Cellulose*, **1**(3), 169-196.

Henrissat, B., Driguez, H., Viet, C., Schülein, M. 1985. Synergism of Cellulases From *Trichoderma reesei* In the Degradation of Cellulose. *Bio/Technology*, **3**, 722-726.

Hermans, P.H., Weidinger, A. 1948. Quantitative X-Ray Investigations on the Crystallinity of Cellulose Fibers. *J. Appl. Phys.*, **19**, 491-506.

Himmel, M.E., Ding, S.-Y., Johnson, D.K., Adney, W.S., Nimlos, M.R., Brady, J.W., Foust, T.D. 2007. Biomass Recalcitrance: Engineering Plants and Enzymes for Biofuels Production. *Science*, **315**, 804-807.

Holtzapple, M., Cognata, M., Shu, Y., Hendrickson, C. 1990. Inhibition of *Trichoderma reesei* Celulase by Sugars and Solvents. *Biotechnol. Bioeng.*, **36**, 275-287.

Holtzapple, M.T., Caram, H.S., Humphrey, A.E. 1984. The HCH-1 model of Enzymatic Cellulose Hydrolysis. *Biotechnol. Bioeng.*, **26**, 775-780.

Hong, J., Ye, X., Zhang, Y.-H.P. 2007. Quantitative Determination of Cellulose Accessibility to Cellulase Based on Adsorption of a Nonhydrolytic Fusion Protein Containing CBM and GFP with Its Applications. *Langmuir*, **23**(25), 12535-12540.

Howell, J.A. 1978. Enzyme deactivation during cellulose hydrolysis. *Biotechnol. Bioeng.*, **20**, 847-863.

Howell, J.A., Stuck, J.D. 1975. Kinetics of Solka Floc Cellulose Hydrolysis by *Trichoderma viride* Cellulase. *Biotechnol. Bioeng.*, **17**, 873 - 893.

http://www.nrel.gov/docs/fy11osti/47764.pdf

http://www.cazy.org/CBM1_structure.html

Huang, A.A. 1975. Kinetic Studies on Insoluble Cellulose-Cellulase System. *Biotechnol. Bioeng.*, **17**, 1421-1433.

Huang, X., Penner, M.H. 1991. Apparent Substrate Inhibition of the *Trichoderma reesei* Cellulase System. *J. Agric. Food Chem.*, **39**, 2096-2100.

Igarashi, K., Koivula, A., Wada, M., Kimura, S., Penttilä, M., Samejima, M. 2009. High Speed Atomic Force Microscopy Visualizes Processive Movement of *Trichoderma reesei* Cellobiohydrolase I on Crystalline Cellulose. *J. Biol. Chem.*, **284**, 36186-36190.

Igarashi, K., Wada, M., Hori, R., Samejima, M. 2006. Surface density of cellobiohydrolase on crystalline celluloses A critical parameter to evaluate enzymatic kinetics at a solid–liquid interface. *FEBS J.*, **273**, 2869-2878.

Jalak, J., Valjamae, P. 2010. Mechanism of Initial Rapid Rate Retardation in Cellobiohydrolase Catalyzed Cellulose Hydrolysis. *Biotechnol. Bioeng.*, **106**(6), 871-883.

Jeoh, T., Ishizawa, C.I., Davis, M.F., Himmel, M., Adney, W.S., Johnson, D.K. 2007. Cellulase Digestibility of Pretreated Biomass Is Limited by Cellulase Accessibility *Biotechnol. Bioeng.*, **98**(1), 112-122.

Jeoh, T., Wilson, D.B., Walker, L.P. 2006. Effect of Cellulase Mole Fraction and Cellulose Recalcitrance on Synergism in Cellulose Hydrolysis and Binding. *Biotechnol. Prog.*, **22**, 270-277.

Jervis, E.J., Haynes, C.A., Kilburn, D.G. 1997. Surface Diffusion of Cellulases and Their Isolated Binding Domains on Cellulose. *J. Biol. Chem.*, **272**(38), 24016-24023.

Jolliffe, I.T. 2002. *Principal Component Analysis*. *2nd ed*. Springer-Verlag New York, Inc. .

Kadam, K.L., Rydholm, E.C., McMillan, J.D. 2004. Development and Validation of a Kinetic Model for Enzymatic Saccharification of Lignocellulosic Biomass. *Biotechnol. Prog.*, **20**, 698-705.

Kim, D.W., Jeong, Y.K., Lee, J.K. 1994. Adsorption kinetics of exoglucanase in combination with endoglucanase from *Trichoderma viride* on microcrystalline cellulose and its influence on synergistic degradation. *Enzyme Microb. Technol.*, **16**, 649-658.

Kim, J.K., Oh, B.R., Shin, H.-J., Eom, C.-Y., Kim, S.W. 2008. Statistical optimization of enzymatic saccharification and ethanol fermentation using food waste. *Process Biochem.*, **43**, 1308-1312.

Kim, S., Holtzapple, M.T. 2006. Effect of structural features enzyme digestibility of corn stover. *Bioresour. Technol.*, **97**, 583-591.

Kim, T.H., Lee, Y.Y. 2005. Pretreatment and fractionation of corn stover by ammonia recycle percolation process. *Bioresour. Technol.*, **96**(18), 2007-2013.

Kipper, K., Väljamäe, P., Johansson, G. 2005. Processive action of cellobiohydrolase Cel7A from Trichoderma reesei is revealed as 'burst' kinetics on fluorescent polymeric model substrates. *Biochem. J.*, **385**, 527-535.

Kleman-Leyer, K.M., Gilkes, N.R., Jr., R.C.M., Kirk, T.K. 1994. Changes in the molecular-size distribution of insoluble celluloses by the action of recombinant *Cellulomonas fimi* cellulases. *Biochem. J.*, **302**, 463-469.

Kleman-Leyer, K.M., Siika-Aho, M., Teeri, T.T., Kirk, T.K. 1996. The Cellulases Endoglucanase I and Cellobiohydrolase II of *Trichoderma reesei* Act Synergistically To Solubilize Native Cotton Cellulose but Not to Decrease Its Molecular Size. *Appl. Environ. Microbiol.*, **62**(8), 2883-2887.

Kongruang, S., Han, M.J., Breton, C.I.G., Penner, M.H. 2004. Quantitative analysis of cellulose reducing ends. *Appl. Biochem. Biotech.*, **113-116**, 213-231.

Kopelman, R. 1988. Fractal Reaction Kinetics. *Science*, **241**, 1620-1626.

Kopelman, R. 1986. Rate Processes on Fractals: Theory, Simulations and Experiments. *J. Stat. Phys.*, **42**(1/2), 185-200.

Koullas, D.P., Christakopoulos, P., Kekos, D., Macris, B.J., Koukios, E.G. 1992. Correlating the Effect of Pretreatment on the Enzymatic Hydrolysis of Straw. *Biotechnol. Bioeng.*, **39**, 113-116.

Krassig, H.A. 1993. *Cellulose: Structure, Accessibility and Reactivity*. Gordon and Breach Science Publishers.

Kumar, R., Wyman, C.E. 2009. Does change in accessibility with conversion depend on both the substrate and pretreatment technology? *Bioresour. Technol.*, **100**(18), 4193-4202.

Kumar, V., Kothari, S., Banker, G.S. 2001. Effect of the Agitation Rate on the Generation of Low-Crystallinity Cellulose from Phosphoric Acid. *J. Appl. Polym. Sci.*, **82**, 2624-2628.

Kurakake, M., Shirasawa, T., Ooshima, H., Converse, A.O., Kato, J. 1995. An Extension of the Harano-Ooshima Rate Expression for Enzymatic Hydrolysis of Cellulose to Account for Changes in the Amount of Adsorbed Cellulase. *Appl. Biochem. Biotech.*, **50**, 231-241.

Kurasin, M., Valjamae, P. 2011. Processivity of Cellobiohydrolases Is Limited by the Substrate. *J. Biol. Chem.*, **286**(1), 166-177.

Kwok, J.T.-Y., Tsang, I.W.-H. 2004. The Pre-Image Problem in Kernel Methods. *IEEE Trans. Neur. Net.*, **15**(6), 1517-1525.

Laidler, K.J. 1955. Theory of the transient phase in kinetics, with special reference to enzyme systems. *Can. J. Chem.*, **33**, 1614-1624.

Laureano-Perez, L., Teymouri, F., Alizadeh, H., Dale, B.E. 2005. Understanding Factors that Limit Enzymatic Hydrolysis of Biomass. *Appl. Biochem. Biotech.*, **121-124**, 1081-1099.

Lee, D.D., Seung, H.S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788-791.

Lee, Y.-H., Fan, L.T., Fan, L.-S. 1980. Kinetics of hydrolysis of insoluble cellulose by cellulase. *Advances in Biochemical Engineering*, **17**, 131-168.

Lee, Y.H., Fan, L.T. 1983. KINETIC-STUDIES OF ENZYMATIC-HYDROLYSIS OF INSOLUBLE CELLULOSE .2. ANALYSIS OF EXTENDED HYDROLYSIS TIMES. *Biotechnol. Bioeng.*, **25**(4), 939-966.

Lenz, J., Esterbauer, H., Sattler, W., Schurz, J., Wrentschur, E. 1990. Changes of structure and morphology of regenerated cellulose caused by acid and enzymatic hydrolysis. *J. Appl. Polym. Sci.*, **41**, 1315-1326.

Levenspiel, O. 1999. *Chemical Reaction Engineering*. John Wiley & Sons.

Levine, S.E., Fox, J.M., Blanch, H.W., Clark, D.S. 2010. A Mechanistic Model of the Enzymatic Hydrolysis of Cellulose. *Biotechnol. Bioeng.*, **107**(1), 37-51.

Liao, J., Warmuth, M.K., Govindarajan, S., Ness, J.E., Wang, R.P., Gustafsson, C., Minshull, J. 2007. Engineering proteinase K using machine learning and synthetic genes. *BMC Biotechnol.*, **7**, 19.

Liao, W., Liu, Y., Wen, Z., Frear, C., Chen, S. 2008. Kinetic Modeling of Enzymatic Hydrolysis of Cellulose in Differently Pretreated Fibers From Dairy Manure. *Biotechnol. Bioeng.*, **101**(3), 441-451.

Liaw, E.-T., Penner, M.H. 1990. Substrate-Velocity Relationships for the Trichoderma viride Cellulase-Catalyzed Hydrolysis of Cellulose. *Appl. Environ. Microbiol.*, **56**(8), 2311-2318.

Lin, Jian-Qiang, Lee, S.-M., Koo, Y.-M. 2005. Modeling and Simulation of Simultaneous Saccharification and Fermentation of Paper Mill Sludge to Lactic Acid. *J. Microbiol. Biotechnol.*, **15**(1), 40-47.

Linder, M., Lindeberg, G., Reinikainen, T., Teeri, T.T., Pettersson, G. 1995a. THE DIFFERENCE IN AFFINITY BETWEEN 2 FUNGAL CELLULOSE-BINDING DOMAINS IS DOMINATED BY A SINGLE AMINO-ACID SUBSTITUTION. *FEBS Lett.*, **372**(1), 96-98.

Linder, M., Mattinen, M.L., Kontteli, M., Lindeberg, G., Stahlberg, J., Drakenberg, T., Reinikainen, T., Pettersson, G., Annila, A. 1995b. IDENTIFICATION OF FUNCTIONALLY IMPORTANT AMINO-ACIDS IN THE CELLULOSE-BINDING DOMAIN OF TRICHODERMA-REESEI CELLOBIOHYDROLASE-I. *Protein Sci.*, **4**(6), 1056-1064.

Lineweaver, H., Burk, D. 1934. The Determination of Enzyme Dissociation Constants. *J. Am. Chem. Soc.*, **56**(3), 658-666.

Ljunggren, M. 2005. Kinetic Analysis and modeling of enzymatic hydrolysis and SSF, Vol. 2005.

Luo, J., Xia, L., Lin, J., Cen, P. 1997. Kinetics of Simultaneous Saccharification and Lactic Acid Fermentation Processes. *Biotechnol. Prog.*, **13**, 762-767.

Lynd, L.R., Laser, M.S., Bransby, D., Dale, B.E., Davison, B., Hamilton, R., Himmel, M., Keller, M., McMillan, J.D., Sheehan, J., Wyman, C.E. 2008. How biotech can transform biofuels. *Nat. Biotechnol.*, **26**, 169-172.

Lynd, L.R., Weimer, P.J., Zyl, W.H.v., Pretorius, I.S. 2002. Microbial Cellulose Utilization: Fundamentals and Biotechnology. *Microbiol. Mol. Biol. Rev.*, **66**(3), 506-577.

Maguire, R.J. 1977. Kinetics of the hydrolysis of cellulose by beta-1,4-glucan cellobiohydrolase of *Trichoderma viride*. *Can. J. Biochem*, **55**, 644-650.

Majdanac, L.D., Poleti, D., Teodorovic, M.J. 1991. Determination of the crystallinity of cellulose samples by X-ray diffraction. *Acta Polym.*, **42**(8), 351-357.

Mansfield, S.D., Meder, R. 2003. Cellulose hydrolysis - the role of monocomponent cellulases in crystalline cellulose degradation. *Cellulose*, **10**, 159-169.

Mansfield, S.D., Mooney, C., Saddler, J.N. 1999. Substrate and Enzyme Characteristics that Limit Cellulose Hydrolysis. *Biotechnol. Prog.*, **15**, 804-816.

Marson, G.A., Seoud, O.A.E. 1999. Cellulose Dissolution in Lithium Chloride/ N,NDimethylacetamide Solvent System: Relevance of Kinetics of Decrystallization to Cellulose Derivatization Under Homogeneous Solution Conditions. *J. Polym. Sci., Part A: Polym. Chem.*, **37**, 3738-3744.

Medve, J., Karlsson, J., Lee, D., Tjerneld, F. 1998. Hydrolysis of Microcrystalline Cellulose by Cellobiohydrolase I and Endoglucanase II from *Trichoderma reesei*: Adsorption, Sugar Production, and Synergism of the Enzymes. *Biotechnol. Bioeng.*, **59**(5), 621-634.

Medve, J., Ståhlberg, J., Tjerneld, F. 1994. Adsorption and Synergism of Cellobiohydrolase I and II of *Trichoderma reesei* During Hydrolysis of Microcrystalline Cellulose. *Biotechnol. Bioeng.*, **44**, 1064-1073.

Medve, J., Ståhlberg, J., Tjerneld, F. 1997. Isotherms for adsorption of cellobiohydrolase I and II from trichoderma reesei on microcrystalline cellulose *Appl. Biochem. Biotech.*, **66**, 39-56.

Meinke, A., Gilkes, N.R., Kilburn, D.G., Jr., R.C.M., Warren, R.A.J. 1993. Cellulose-Binding Polypeptides from *Cellulomonas fim*i: Endoglucanase D (CenD), a Family A b-1,4-Glucanase. *J. Bacteriol.*, **175**(7), 1910-1918.

Meyer, M.M., Silberg, J.J., Voigt, C.A., Endelman, J.B., Mayo, S.L., Wang, Z.G., Arnold, F.H. 2003. Library analysis of SCHEMA-guided protein recombination. *Protein Sci.*, **12**(8), 1686-1693.

Michaelis, L., Menten, M.L. 1913. Die kinetik der invertinwirkung. *Biochem. Z.*, **49**, 333-369.

Moldes, A.B., Alonso, J.L., Parajó, J.C. 1999. Cogeneration of Cellobiose and Glucose from Pretreated Wood and Bioconversion to Lactic Acid: A Kinetic Study. *J. Biosci. Bioeng.*, **87**(6), 787-792.

Moon, H., Kim, J.-S., Oh, K.-K., Kim, S.-W., Hong, S.-I. 2001. Kinetic Modeling of Simultaneous Saccharification and Fermentation for Ethanol Production Using Steam-Exploded Wood with Glucose- and Cellobiose-Fermenting Yeast, *Brettanomyces custersii*. *J. Microbiol. Biotechnol.*, **11**(4), 598-606.

Movagarnejad, K., Sohrabi, M., Kaghazchi, T., Vahabzadeh, F. 2000. A model for the rate of enzymatic hydrolysis of cellulose in heterogeneous solid–liquid systems. *Biochem. Engg. J.*, **4**, 197-206.

Movagharnejad, K. 2005. Modified shrinking particle model for the rate of enzymatic hydrolysis of impure cellulosic waste materials with enzyme reuse by the substrate replacement. *Biochem. Engg. J.*, **24**, 217-223.

Movagharnejad, K., Sohrabi, M. 2003. A model for the rate of enzymatic hydrolysis of some cellulosic waste materials in heterogeneous solid–liquid systems. *Biochem. Engg. J.*, **14**, 1-8.

Mulakala, C., Reilly, P.J. 2005. Hypocrea jecorina (Trichoderma reesei) Cel7A as a Molecular Machine: A Docking Study. *Proteins: Struct., Funct., Bioinf.*, **60**, 598-605.

Nakai, Y., Fukuoka, E., Nakajima, S., Hasegawa, J. 1977. Crystallinity and Physical Characteristics of Microcrystalline Cellulose. *Chem. Pharm. Bull.*, **25**, 96-101.

Nakasaki, K., Murai, T., Akiyama, T. 1988. Kinetic Modeling of Simultaneous Saccharification and Fermentation of Cellulose. *J. Chem. Eng. Jpn.*, **21**(4), 436-438.

Nassar, R., Chou, S.T., Fan, L.T. 1991. Stochastic Anaysis of Stepwise Cellulose Degradation. *Chem. Eng. Sci.*, **46**(7), 1651-1657.

Nidetzky, B., Steiner, W. 1993. A New Approach for Modeling Cellulase-Cellulose Adsorption and the Kinetics of the Enzymatic Hydrolysis of Microcrystalline Cellulos. *Biotechnol. Bioeng.*, **42**, 469-479.

Nidetzky, B., Steiner, W., Claeyssens, M. 1994a. Cellulose hydrolysis by the cellulases from *Trichoderma reesei*: adsorptions of two cellobiohydrolases, two endocellulases and their core proteins on filter paper and their relation to hydrolysis. *Biochem. J.*, **308**, 817-823.

Nidetzky, B., Steiner, W., Hayn, M., Claeyssens, M. 1994b. Cellulose hydrolysis by the cellulases from *Trichoderma reesei*: a new model for synergistic interaction. *Biochem. J.*, **298**, 705-710.

Nidetzky, B., Steiner, W., Hayn, M., Esterbauer, H. 1993. Enzymatic Hydrolysis of Wheat Straw After Steam Pretreatment: Experimental Data and Kinetic Modeling. *Bioresour. Technol.*, **44**, 25-32.

Nidetzky, B., Zachariae, W., Gercken, G., Hayn, M., Steiner, W. 1994c. Hydrolysis of cellooligosaccharides by *Trichoderma ressei* cellobiohydrolases: Experimental data and kinetic modeling. *Enzyme Microb. Technol.*, **16**, 43-52.

Nishiyama, Y., Langan, P., Chanzy, H. 2002. Crystal structure and hydrogen-bonding system in cellulose Ib from synchroton X-ray and neutron fiber diffraction. *J. Am. Chem. Soc.*, **124**, 9074-9082.

Nov, Y., Wein, L.M. 2005. Modeling and analysis of protein design under resource constraints. *J. Comput. Biol.*, **12**(2), 247-282.

Nutor, J.R.K., Converse, A.O. 1991. The Effect of Enzyme and Substrate Levels on the Specific Hydrolysis Rate of Pretreated Poplar Wood. *Appl. Biochem. Biotech.*, **28/29**, 757-771.

O'Dwyer, J.P., Zhu, L., Granda, C.B., Chang, V.S., Holtzapple, M.T. 2008. Neural Network Prediction of Biomass Digestibility Based on Structural Features. *Biotechnol. Prog.*, **24**, 283-292.

O'Dwyer, J.P., Zhu, L., Granda, C.B., Holtzapple, M.T. 2007. Enzymatic hydrolysis of lime-pretreated corn stover and investigation of the HCH-1 Model: Inhibition pattern, degree of inhibition, validity of simplified HCH-1 Model. *Bioresour. Technol.*, **98**, 2969-2977.

Oh, K.-K., Kim, S.-W., Jeong, Y.-S., Hong, S.-I. 2000. Bioconversion of Cellulose into Ethanol by Nonisothermal Simultaneous Saccharification and Fermentation. *Appl. Biochem. Biotech.*, **89**, 15-30.

Ohmine, K., Ooshima, H., Harano, Y. 1983. Kinetic Study on Enzymatic Hydrolysis of Cellulose by Cellulase from *Trichoderma Viride*. *Biotechnol. Bioeng.*, **25**, 2041-2053.

Okazaki, M., Moo-Young, M. 1978. Kinetics of Enzymatic Hydrolysis of Cellulose: Analytical description of a mechanistic model. *Biotechnol. Bioeng.*, **20**, 637-663.

Ooshima, H., Kurakake, M., Kato, J., Harano, Y. 1991. Enzymatic Activity of Cellulose Adsorbed on Cellulose and its change during Hydrolysis. *Appl. Biochem. Biotech.*, **31**, 253-266.

Ooshima, H., Sakata, M., Harano, Y. 1983. Adsorption of Cellulase from *Trichoderma viride* on Cellulose. *Biotechnol. Bioeng.*, **25**, 3103-3114.

Pala, H., Mota, M., Gama, F.M. 2007. Enzymatic depolymerisation of cellulose. *Carbohydr. Polym.*, **68**, 101-108.

Parajó, J.C., Alonso, J.L., Santos, V. 1996. Development of a Generalized Phenomenological Model Describing the Kinetics of the Enzymatic Hydrolysis of NaOH-Treated Pine Wood. *Appl. Biochem. Biotech.*, **56**, 289-299.

Park, E.Y., Ikeda, Y., Okuda, N. 2002. Empirical Evaluation of Cellulase on Enzymatic Hydrolysis of Waste Office Paper. *Biotechnol. Bioprocess Eng.*, **7**, 268-274.

Park, S., Baker, J.O., Himmel, M.E., Parilla, P.A., Johnson, D.K. 2010. Cellulose crystallinity index: measurement techniques and their impact on interpreting cellulase performance. *Biotechnol. Biofuels*, **3**(10).

Park, S., Johnson, D.K., Ishizawa, C.I., Parilla, P.A., Davis, M.F. 2009. Measuring the crystallinity index of cellulose by solid state ^{13}C nuclear magnetic resonance. *Cellulose*, **16**, 641-647.

Peiterson, N., Edward W. Ross, J. 1979. Mathematical Model for Enzymatic Hydrolysis and Fermentation of Cellulose by *Trichoderma*. *Biotechnol. Bioeng.*, **997-1017**.

Peri, S., Karra, S., Lee, Y.Y., Karim, M.N. 2007. Modeling Intrinsic Kinetics of Enzymatic Cellulose Hydrolysis. *Biotechnol. Prog.*, **23**, 626-637.

Pettersson, P.O., Eklund, R., Zacchi, G. 2002. Modeling Simultaneous Saccharification and Fermentation of Softwood. *Appl. Biochem. Biotech.*, **98-100**, 733-746.

Philippidis, G.P., Smith, T.K., Wyman, C.E. 1993. Study of the Enzymatic Hydrolysis of Cellulose for Production of Fuel Ethanol by the Simultaneous Saccharification and Fermentation Process. *Biotechnol. Bioeng.*, **41**, 846-853.

Philippidis, G.P., Spindler, D.D., Wyman, C.E. 1992. Mathematical Modeling of Cellulose Conversion to Ethanol by the Simultaneous Saccharification and Fermentation Process. *Appl. Biochem. Biotech.*, **34/35**, 543-556.

Podkaminer, K.K., Shao, X.J., Hogsett, D.A., Lynd, L.R. 2011. Enzyme Inactivation by Ethanol and Development of a Kinetic Model for Thermophilic Simultaneous Saccharification and Fermentation at 50 degrees C with Thermoanaerobacterium saccharolyticum ALK2. *Biotechnol. Bioeng.*, **108**(6), 1268-1278.

Polizzi, S., Fagherazzi, G., Benedetti, A., Battagliarin, M. 1990. A Fitting Method for the Determination of Crystallinity by Means of X-ray Diffraction. *J. Appl. Crystallogr.*, **23**, 359-365.

Praestgaard, E., Elmerdahl, J., Murphy, L., Nymand, S., McFarland, K.C., Borch, K., Westh, P. 2011. A kinetic model for the burst phase of processive cellulases. *FEBS J.*, **278**(9), 1547-1560.

Puls, J., Wood, T.M. 1991. The Degradation Pattern of Cellulose by Extracellular Cellulases of Aerobic and Anaerobic Microorganisms. *Bioresour. Technol.*, **36**, 15-19.

Rabinovich, M.L., Melnick, M.S., Bolobova, A.V. 2002. The Structure and Mechanism of Action of Cellulolytic Enzymes. *Biochemistry (Moscow)*, **67**(8), 850-871.

Rammos, L.P., Nazhad, M.M., Saddler, J.N. 1993. Effect of enzymatic hydrolysis on the morphology and fine structure of pretreated cellulosic residues. *Enzyme Microb. Technol.*, **15**, 821-831.

Ramos, L.A., Assaf, J.M., Seoud, O.A.E., Frollini, E. 2005. Influence of the Supramolecular Structure and Physicochemical Properties of Cellulose on Its Dissolution in a Lithium Chloride/N,N-Dimethylacetamide Solvent System. *Biomarcomolecules*, **6**, 2638-2647.

Reetz, M.T., Carballeira, J.D. 2007. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat. Protoc.*, **2**(4), 891-903.

Rietveld, H.M. 1969. A Profile Refinement Method for Nuclear and Magnetic Structures. *J. Appl. Crystallogr.*, **2**, 65-71.

Ruland, W. 1961. X-ray Determination of Crystallinity and Diffuse Disorder Scattering. *Acta Crystallogr.*, **14**, 1180-1185.

Ryu, D.D.Y., Lee, S.B., Tassinari, T., Macy, C. 1982. Effect of Compression Milling on Cellulose Structure and on Enzymatic Hydrolysis Kinetics. *Biotechnol. Bioeng.*, **24**, 1047-1067.

Sattler, W., Esterbauer, H., Glatter, O., Steiner, W. 1989. The Effect Enzyme Concentration on the Rate of the Hydrolysis of Cellulose. *Biotechnol. Bioeng.*, **33**, 1221-1234.

Savageau, M.A. 1995. Michaelis-Menten Mechanism Reconsidered: Implications of Fractal Kinetics. *J. Theor. Biol.*, **176**, 115-124.

Scheiding, W., Thoma, M., Ross, A., Schugerl, K. 1984. Modelling of the enzymatic hydrolysis of cellobiose and cellulose by a complex enzyme mixture of *Trichoderma reesei* QM9414. *Appl. Microbiol. Biotechnol.*, **20**, 176-182.

Schell, D.J., Ruth, M.F., Tucker, M.P. 1999. Modeling the Enzymatic Hydrolysis of Dilute-Acid Pretreated Douglas Fir. *Appl. Biochem. Biotech.*, **77-79**, 67-81.

Schenzel, K., Fischer, S., Brendler, E. 2005. New method for determining the degree of cellulose I crystallinity by means of FT Raman spectroscopy. *Cellulose*, **12**, 223-231.

Schmid, G., Wandrey, C. 1989. Characterization of a cellodextrin glucohydrolase with soluble oligomeric substrate: experimental results and modeling concentration-time-course data. *Biotechnol. Bioeng.*, **33**, 1445-1460.

Schnell, S. 2003. A Century of Enzyme Kinetics: Reliability of the K_M and v_{max} Estimates. *Comments on Theoretical Biology*, **8**, 169-187.

Schubert, C. 2006. Can biofuels finally take center stage? *Nat. Biotechnol.*, **24**, 777-784.

Schulein, M. 2000. Protein engineering of cellulases. *Biochim. Biophys. Acta*, **1543**, 239-252.

Schurz, J., Janosi, A., Zipper, P. 1987. Röntgenographische Kristallinitätsuntersuchungen an Zallstoffen. *Das Papier*, **41**, 673-679.

Schurz, J., Klapp, H. 1976. Untersuchungen an mikrokristallinen und mikrofeinen Cellulosen. *Papier*, **30**, 510-513.

Segal, L., Creely, J.J., Martin, A.E., Conrad, C.M. 1959. An Empirical Method for Estimating the Degree of Crystallinity of Native Cellulose Using the X-Ray Diffractometer. *Text. Res. J.*, **29**(10), 786-794.

Shao, X., Lynd, L., Wyman, C. 2009a. Kinetic Modeling of Cellulosic Biomass to Ethanol Via Simultaneous Saccharification and Fermentation: Part II. Experimental Validation Using Waste Paper Sludge and Anticipation of CFD Analysis. *Biotechnol. Bioeng.*, **102**(1), 66-72.

Shao, X., Lynd, L., Wyman, C., Bakker, A. 2009b. Kinetic Modeling of Cellulosic Biomass to Ethanol Via Simultaneous Saccharification and Fermentation: Part I. Accomodation of Intermittent Feeding and Analysis of Staged Reactors. *Biotechnol. Bioeng.*, **102**(1), 59-65.

Shen, J., Agblevor, F.A. 2008a. Kinetics of Enzymatic Hydrolysis of Steam-Exploded Cotton Gin Waste. *Chem. Eng. Commun.*, **195**(9), 1107-1121.

Shen, J., Agblevor, F.A. 2008b. Optimization of enzyme loading and hydrolytic time in the hydrolysis of mixtures of cotton gin waste and recycled paper sludge for the maximum profit rate. *Biochem. Engg. J.*, **41**, 241-250.

Shen, Y., Wang, L.M., Sun, K. 2004. Kinetics of the Cellulase Catalyzed Hydrolysis of Cellulose Fibres. *Text. Res. J.*, **74**, 539-545.

Shin, Donggyun, Yoo, A., Kim, S.W., Yang, D.R. 2006. Cybernetic Modeling of Simultaneous Saccharification and Fermentation for Ethanol Production from Steam-Exploded Wood with *Brettanomyces custersii*. *J. Microbiol. Biotechnol.*, **16**(9), 1355-1361.

Siegel, J.B., Zanghellini, A., Lovick, H.M., Kiss, G., Lambert, A.R., Clair, J.L.S., Gallaher, J.L., Hilvert, D., Gelb, M.H., Stoddard, B.L., Houk, K.N., Michael, F.E., Baker, D. 2010. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science*, **329**(5989), 309-313.

Sin, G., Meyer, A.S., Gernaey, K.V. 2009. Are mechanistic cellulose-hydrolysis models reliable for use in biofuel process design? – Identifiability and sensitivity analysis. in: *Proceedings of the 7th international conference on the foundations of computer-aided process design (FOCAPD).*

Soltys, J., Lisowski, Z., Knapczyk, J. 1984. X-Ray Diffraction Study of the Crystallinity Index and the Structure of the Microcrystalline Cellulose. *Acta Pharmaceutica Technologica*, **30**(2), 174-180.

South, C.R., Hogsett, D.A.L., Lynd, L.R. 1995. Modeling simultaneous saccharification and fermentation of lignocellulose to ethanol in batch and continuous reactors. *Enzyme Microb. Technol.*, **17**, 797-803.

Souza, I.J.D., Bouchard, J., Methot, M., Berry, R., Argyropoulos, D.S. 2002. Carbohydrates in Oxygen Delignification. Part I: Changes in Cellulose Crystallinity. *J. Pulp Pap. Sci.*, **28**(5), 167-170.

Ståhlberg, J., Johannson, G., Pettersson, G. 1991. A New Model for Enzymatic Hydrolysis of Cellulose Based on the Two-Domain Structure of Cellobiohydrolase I. *Nat. Biotechnol.*, **9**, 286-290.

Steiner, W., Sattler, W., Esterbauer, H. 1988. Adsorption of *Trichoderma reesei* Cellulase on Cellulose: Experimental Data and Their Analysis by Different Equations. *Biotechnol. Bioeng.*, **32**, 853-865.

Steipe, B., Schiller, B., Pluckthun, A., Steinbacher, S. 1994. Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. *J. Mol. Biol.*, **240**(3), 188-192.

Stemmer, W.P.C. 1994. RAPID EVOLUTION OF A PROTEIN IN-VITRO BY DNA SHUFFLING. *Nature*, **370**(6488), 389-391.

Stenberg, K., Bollók, M., Réczey, K., Galbe, M., Zacchi, G. 2000. Effect of Substrate and Cellulase Concentration on Simultaneous Saccharification and Fermentation of Steam-Pretreated Softwood for Ethanol Production. *Biotechnol. Bioeng.*, **68**(2), 204-210.

Suga, K., Dedem, G.v., Moo-Young, M. 1975. Degradation of Polysaccharides by Endo and Exo Enzymes: A Theoretical Analysis. *Biotechnol. Bioeng.*, **17**, 433-439.

Sun, Y., Cheng, J. 2002. Hydrolysis of lignocellulosic materials for ethanol production: a review. *Bioresour. Technol.*, **83**, 1-11.

Szijártó, N., Siika-aho, M., Tenkanen, M., Alapuranen, M., Vehmaanperä, J., Réczey, K., Viikari, L. 2008. Hydrolysis of amorphous and crystalline cellulose by heterologously produced cellulases of *Melanocarpus albomyces*. *J. Biotechnol.*, **136**, 140-147.

Tarantili, P.A., Koullas, D.P., Christakopoulos, P., Kekos, D., Koukios, E.G., Macris, B.J. 1996. Cross-Synergism in Enzymatic Hydrolysis of Lignocellulosics: Mathematical Correlations According to a Hyperbolic Model. *Biomass and Bioenergy*, **10**(4), 213-219.

Teeäär, R., Serimaa, R., Paakkari, T. 1987. Crystallinity of cellulose, as determined by CP/MAS NMR and XRD methods. *Polym. Bull.*, **17**, 231-237.

Teeri, T.T. 1997. Crystalline cellulose degradation: new insight into the function of cellobiohydrolases. *Trends Biotechnol.*, **15**, 160-167.

Tenenbaum, J.B., Silva, V.d., Langford, J.C. 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, **290**, 2319-2323.

The Mathworks Inc. R2008b. MATLAB.

The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC.

Thomas, J., Ramakrishnan, N., Bailey-Kellogg, C. 2008. Graphical models of residue coupling in protein families. *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, **5**(2), 183-197.

Thomas, J., Ramakrishnan, N., Bailey-Kellogg, C. 2009. Protein Design by Sampling an Undirected Graphical Model of Residue Constraints. *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, **6**(3), 506-516.

Thompson, J.D., Higgins, D.G., Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673-4680.

Thygesen, A., Oddershede, J., Lilholt, H., Thomsen, A.B., Ståhl, K. 2005. On the determination of crystallinity and cellulose content in plan fibres. *Cellulose*, **12**, 563-576.

Ting, C.L., Marakov, D.E., Wang, Z.-G. 2009. A Kinetic Model for the Enzymatic action of Cellulase. *J. Phys. Chem. B*, **113**(14), 4970-4977.

Tu, M., Chandra, R.P., Saddler, J.N. 2007. Evaluating the Distribution of Cellulases and the Recycling of Free Cellulases during the Hydrolysis of Lignocellulosic Substrates. *Biotechnol. Prog.*, **23**, 398-406.

Väljamäe, P., Kipper, K., Petterson, G., Johansson, G. 2003. Synergistic Cellulose Hydrolysis Can Be Described in Terms of Fractal-Like Kinetics. *Biotechnol. Bioeng.*, **84**(2), 254-257.

Väljamäe, P., Petterson, G., Johansson, G. 2001. Mechanism of substrate inhibition in cellulose synergistic degradation. *Eur. J. Biochem.*, **268**, 4520-4526.

Väljamäe, P., Sild, V., Nutt, A., Pettersson, G., Johansson, G. 1999. Acid hydrolysis of bacterial cellulose reveals different modes of synergistic action between cellobiohydrolase I and endoglucanase I. *Eur. J. Biochem.*, **266**, 327-334.

Väljamäe, P., Sild, V., Pettersson, G., Johansson, G. 1998. The initial kinetics of hydrolysis by cellobiohydrolases I and II is consistent with a cellulose surface - erosion model. *Eur. J. Biochem.*, **253**, 469-475.

Vásquez, M.P., Silva, J.N.C.d., Jr., M.B.d.S., Jr., N.P. 2007. Enzymatic Hydrolysis Optimization to Ethanol Production by Simultaneous Saccharification and Fermentation. *Appl. Biochem. Biotech.*, **136-140**, 141-153.

Vonk, C.G. 1973. Computerization of Ruland's X-ray Method for Determination of the Crystallinity in Polymers. *J. Appl. Crystallogr.*, **6**, 148-152.

Wakelin, J.H., Virgin, H.S., Crystal, E. 1959. Development and Comparison of Two X-Ray Methods for Determining the Crystallinity of Cotton Cellulose. *J. Appl. Phys.*, **30**(11), 1654-1662.

Wald, S., Wilke, C.R., Blanch, H.W. 1984. Kinetics of the Enzymatic Hydrolysis of Cellulose. *Biotechnol. Bioeng.*, **26**, 221-230.

Wang, M., Yang, J., Xu, Z.-J., Chou, K.-C. 2005. SLLE for predicting membrane protein types. *J. Theor. Biol.*, **232**, 7-15.

Wang, Z.L., Feng, H. 2010. Fractal kinetic analysis of the enzymatic saccharification of cellulose under different conditions. *Bioresour. Technol.*, **101**(20), 7995-8000.

Wood, T.M. 1975. Properties and modes of action of cellulases. *Biotechnology and Bioengineering Symposium*, **5**, 111-137.

Wood, T.M., McCrae, S.I. 1978. The cellulase of Trichoderma koningii. Purification and properties of some endoglucanase components with special reference to their action on cellulose when acting alone and in synergism with the cellobiohydrolase. *Biochem. J.*, **171**, 61-72.

Woodward, J., Hayes, M.K., Lee, N.E. 1988a. Hydrolysis of Cellulose by Saturating and Non−Saturating Concentrations of Cellulase: Implications for Synergism. *Bio/Technology*, **6**, 301-304.

Woodward, J., Lima, M., Lee, N.E. 1988b. The role of cellulase concentration in determining the degree of synergism in the hydrolysis of microcrystalline cellulose. *Biochem. J.*, **255**, 895-899.

Wu, J., Ju, L.K. 1998. Enhancing enzymatic saccharification of waste newsprint by surfactant addition. *Biotechnol. Prog.*, **14**(4), 649-652.

Xiao, Z., Zhang, X., Gregg, D.J., Saddler, J.N. 2004. Effects of Sugar Inhibition on Cellulases and Glucosidase During Enzymatic Hydrolysis of Softwood Substrates. *Appl. Environ. Microbiol.*, **113-116**, 1115-1126.

Xu, F., Ding, H. 2007. A new kinetic model for heterogeneous (or spatially confined) enzymatic catalysis: Contributions from the fractal and jamming (overcrowding) effects. *Appl. Catal., A*, **317**, 70-81.

Yang, B., Willies, D.M., Wyman, C.E. 2006. Changes in the Enzymatic Hydrolysis Rate of Avicel Cellulose With Conversion. *Biotechnol. Bioeng.*, **94**(6), 1122-1128.

Yue, Z., Bin, W., Baixu, Y., Peiji, G. 2004. Mechanism of cellobiose inhibition in cellulose hydrolysis by cellobiohydrolase. *Sci. China, Ser. C*, **47**(1), 18-24.

Zhang, S., Wolfgang, D.E., Wilson, D.B. 1999. Substrate Heterogeneity Causes the Nonlinear Kinetics of Insoluble Cellulose Hydrolysis. *Biotechnol. Bioeng.*, **66**(1), 35-41.

Zhang, Y.-H.P., Lynd, L.R. 2005. Determination of the Number-Average Degree of Polymerization of Cellodextrins and Cellulose with Application to Enzymatic Hydrolysis. *Biomacromolecules*, **6**, 1510-1515.

Zhang, Y.-H.P., Lynd, L.R. 2006. A Functionally Based Model for Hydrolysis of Cellulose by Fungal Cellulase. *Biotechnol. Bioeng.*

Zhang, Y.-H.P., Lynd, L.R. 2004. Toward an Aggregated Understanding of Enzymatic Hydrolysis of Cellulose: Noncomplexed Cellulase Systems. *Biotechnol. Bioeng.*, **88**(7), 797-824.

Zhang, Y., Xu, J.L., Xu, H.J., Yuan, Z.H., Guo, Y. 2010. Cellulase deactivation based kinetic modeling of enzymatic hydrolysis of steam-exploded wheat straw. *Bioresour. Technol.*, **101**(21), 8261-8266.

Zheng, Y., Pan, Z., Zhang, R., Jenkins, B.M. 2009. Kinetic Modeling for Enzymatic Hydrolysis of Pretreated Creeping Wild Ryegrass. *Biotechnol. Bioeng.*, **102**(6), 1558-1569.

Zhou, J., Wang, Y.-H., Chu, J., Luo, L.-Z., Zhuang, Y.-P., Zhang, S.-L. 2009a. Optimization of cellulase mixture for efficient hydrolysis of steam-exploded corn stover by statistically designed experiments. *Bioresour. Technol.*, **100**, 819-825.

Zhou, W., Hao, Z., Xu, Y., Schuttler, H.-B. 2009b. Cellulose Hydrolysis in Evolving Substrate Morphologies II: Numerical Results and Analysis. *Biotechnol. Bioeng.*, **104**(2), 275-289.

Zhou, W., Schuttler, H.-B., Hao, Z., Xu, Y. 2009c. Cellulose Hydrolysis in Evolving SUbstrate Morphologies I: A General Modeling Formalism. *Biotechnol. Bioeng.*, **104**(2), 261-274.

Zhu, L., O'Dwyer, J.P., Chang, V.S., Granda, C.B., Holtzapple, M.T. 2008. Structural features affecting biomass enzymatic digestibility. *Bioresour. Technol.*, **99**, 3817-3828.

VITA

PRABUDDHA BANSAL

Prabuddha Bansal was born in Chandigarh, India. He attended St. John's High School in Chandigarh, and received Bachelor of Technology in Chemical Engineering with a minor in Industrial Engineering from IIT Madras, India in 2007 before coming to Georgia Tech to pursue a doctorate in Chemical & Biomolecular Engineering. Prabuddha is proficient in modeling of chemical and biomolecular systems, statistics, data-driven protein engineering, and also has experimental experience in the field of biocatalysis. He has numerous academic achievements to his name – securing high ranks in Physics and Mathematics Olympiads, IIT-JEE entrance test, academic exemplary achievement award and outstanding qualification performance award at Georgia Tech. Outside school, he is an active soccer and squash player. He was a part of the IIT Madras soccer team, and was the captain in 2005. He also represented Georgia Tech in a South-East college squash tournament in February 2011. Recently, in December 2010, Prabuddha successfully climbed the Mt. Kilimanjaro (5895 m/19341 ft. a.s.l).